Cognitive Effort in Post-Editing of Machine Translation: Evidence from Eye Movements, Subjective Ratings, and Think-Aloud Protocols

Lucas Nunes Vieira

Doctor of Philosophy

School of Modern Languages, Newcastle University, UK

December 2015

Abstract

This thesis investigates the expenditure of cognitive effort in post-editing of machine translation. A mixed-method approach involving the use of eye movements, subjective ratings and think-aloud protocols was adopted for the investigation. The project aims at revealing connections between cognitive effort and variables including linguistic characteristics of the source text and the machine-translation output, post-editors' individual traits, different linguistic aspects of the activity attended to during the task, and the overall quality of the post-edited text, assessed by human translators in terms of fluency (linguistic quality) and adequacy (faithfulness to the source text). Two tasks were conducted to pursue these aims: one involving eye tracking and a self-report scale of cognitive effort, and another carried out by a different, but comparable, sample of participants, under a think-aloud condition. Results indicate that variables such as an automatic machine-translation quality score and source-text type-token ratio are good predictors of cognitive effort in post-editing. The relationship between cognitive effort and post-editors' traits was found to be a complex one, with significant links in this respect only appearing in the context of interactions between variables. A complex relationship was also found between editing behaviour and the quality of the post-edited text: the amount of changes implemented was found to have a generally positive association with post-edited fluency, though cognitive effort was found to be negatively correlated with both the fluency and adequacy of the post-edited text. Mental processes involving grammar and lexis were significantly related to the levels of cognitive effort expended by participants, being also the aspects most frequently attended to in the activity. From a methodological perspective, despite the criticisms received by the think-aloud method in previous research, empirical data obtained in this project indicate that think-aloud protocols correlate with eye movements and subjective ratings as measures of cognitive effort.

Acknowledgments

First I would like to thank my main supervisor, Dr Francis Jones, for his attention to detail and the unswerving help and support he provided me with throughout this project – his attentive consideration of my research ideas made this thesis possible. Thanks also go to my second and third supervisors: Dr Michael Jin, for his positive criticism and for useful discussions on statistics and experimental design; and Dr Ya-Yun Chen, for her perceptive suggestions and for being a great mentor in all things academic.

Thanks go to Dr Lucia Specia for useful discussions that helped shape an earlier version of this project; to Dr Sandra Monteiro for her comments on the project proposal; to Dr Peter Flom for his suggestions on the analysis of ordinal data; to Dr Jorge Baptista for introducing me to quantitative methods in linguistics; and to Dr Ceila Maria Ferreira, for stimulating my interest in research early on.

Mãe, pai, Carol e João Gustavo, obrigado pelo apoio, pelas risadas e pela paciência.

Thanks go to Adam for his support, for the best stews England has seen and for reading everything I write and taking issue with my excessive use of words.

To Michael and Ruth for their help with proofreading.

To all those who took part in the research as post-editor, translation assessor, or coder – for lending their minds to this study.

And finally, to the School of Modern Languages at Newcastle University – for funding this project.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BLEU – BiLingual Evaluation Understudy

GTM – General Text Matcher

HTER – Human-Targeted Translation Edit Rate

Meteor – Metric for Evaluation of Translation with Explicit Ordering

MT – machine translation

PC – principal component

PE – post-editing

RQ – research question

SD – standard deviation

ST – source text

TAPs – think-aloud protocols

TAUS – Translation Automation User Society

TER – Translation Edit Rate

TT – target text

WMC – working memory capacity

Chapter 1. Introduction

This thesis is an empirical investigation of effort expenditure in post-editing of machine translation (MT). Normally referred to as just ‘post-editing’ (PE), this activity consists of editing the output of MT systems as opposed to translating texts from scratch. MT is in fact one of many features that have been introduced to the human translation workflow of late as a result of recent advancements in language technology. Another feature of this kind are so-called translation memories, i.e. databases containing aligned strings of source text (ST) and human translation which are searched for ST similarities (between the strings being translated and those stored in the database) that allow previous translations to be recycled. Though MT and translation memories can be combined in providing the first draft of a translation to be edited, in the present thesis and in the vast majority of previous research, post-editing refers strictly to the act of editing translations produced automatically by MT systems, a practice that has introduced a number of changes to the way translation (mostly non-literary) is carried out in certain professional settings.

Since the first attempt at automatically translating texts in the first half of the twentieth century, MT has undergone extremely active periods of research as well as rather inactive ones, when frustrated expectations led to funding cuts in the area. From the beginning of the 1990s until the present moment, however, professional translation is going through a new phase where an increasing number of companies and public bodies are making use of MT systems as an integral part of the human translation workflow. This seems due to the fact that, globally, never before has there been such a large volume of content to be translated (see e.g. FreePressRelease 2011), which makes the possibility of translating texts automatically extremely appealing. In addition, computational resources have improved considerably since the early years of MT research, with fairly high levels of MT quality being achieved, making PE a current market reality (see e.g. TAUS 2010). Much research has been done comparing the levels of human effort, speed and translation quality associated with PE and with traditional (i.e. fully human) translation, with PE proving to be a translation modality that is able to increase translators' productivity as well as improve product quality (e.g. Green, Heer and Manning 2013; Plitt and Masselot 2010).

1.1. Overall Research Purpose

Perhaps due to the economic implications of this practice, it is fair to say that the implementation of PE in professional contexts has forged ahead of an in-depth understanding of the underlying mechanisms of the activity. This calls for research that can be applied to a better-informed implementation of PE as well as to more general issues relating to the interaction between humans and machines in the production of language, such as finding out how this interaction actually comes about and how it shapes human (editing) behaviour. As a way of addressing some of these issues, the overarching purpose of the present thesis is to investigate different potential factors that might relate to the expenditure of cognitive effort in PE whilst also looking at certain methodological aspects pertaining to cognitive-effort estimation.

By drawing upon theories from the areas of cognitive and educational psychology, cognitive effort is regarded in this thesis as the amount of mental resources individuals allocate to a task, which may vary according to demands that are intrinsic to the task as well as to characteristics of individuals themselves (see section 2.2.1.3 for a detailed discussion).

Overall PE effort is deemed to comprise various types of effort, with the cognitive type being regarded as the main one, controlling the overall effort expenditure in the activity (see section 2.2.1.1). Indeed, the fact that *cognitive* effort is estimated in this thesis is a distinguishing feature of the investigation. Mounting evidence from previous studies suggests that aspects such as number of editing operations and editing time are, alone, not reliable indicators of the overall effort that goes into the activity. Koponen (2012), for example, found that post-editors may report expending high levels of effort even when few editing operations are performed. In addition, Koponen, Aziz, Ramos and Specia (2012) showed that different errors may lead to corrections that require equal amounts of editing but different amounts of time, providing further evidence of the pitfalls involved in taking only gross indicators of amount of editing into account. Regarding purely temporal measures, de Almeida (2013) found no correlation between post-editors' level of professional experience and time spent post-editing, which led her to suggest that 'PE effort and PE performance involve a high level of complexity that cannot be explained only by analysing temporal values' (ibid., 199). These findings were taken into account in this thesis in setting out to investigate cognitive effort, rather than PE time or number of editing operations, which are related to cognitive effort but may fail to perfectly reflect it.

Both process- and product-based elements (i.e. elements relating to the act of carrying out the task as well as to characteristics of the resulting product) are investigated here in view of their connection with cognitive effort. Specific variables considered for this purpose include estimated levels of MT-output quality, linguistic features of both the ST and MT output, post-editors' individual traits, the mental processes that take place during the activity, and the quality of the post-edited text. To the author's knowledge, this thesis constitutes the first attempt at jointly analysing these elements in view of their relationship with cognitive effort. Findings in this respect can serve a number of purposes, including predicting the effort posed by the activity, revealing more efficient ways of carrying out PE and helping to further characterise the concept of cognitive effort in this specific context.

1.2. Research Methodology

The topics mentioned above are investigated here based on empirical data collected in the context of two PE tasks carried out by participants with relevant experience (either as students or professionals) in translation, PE, or traditional revision (i.e. revision of human translations). The investigation is carried out based on a mixed-method strategy that combines the use of eye tracking, a self-report scale borrowed from the field of educational psychology and think-aloud protocols (TAPs), i.e. spoken data gathered through a technique that consists of asking research participants to 'think-aloud' as they carry out the task. The analysis of data stemming from these sources rests on a quantitative/statistical backbone which is expanded with qualitative examples and discussion, according both breadth and depth to the study whilst also complying with the principle of triangulation (Creswell 2009), whereby the strengths of different methodologies converge to form a more robust and reliable overall methodological approach to an investigation.

The quantitative component of the analysis relies on inferential mixed-effects regression modelling, a rigorous statistical technique that compensates for effects relating just to the participants or materials sampled for the study, rendering findings more generalizable. Results in this respect serve to shape and inform a more in-depth qualitative analysis of the data which, where appropriate, also counts on reliability cross-checks; for example by including an external researcher in the project to measure data coding reliability.

1.3. Structure of the Thesis

The remainder of this thesis is structured as follows:

Chapter 2 provides a review of relevant literature in translation studies and PE as well as an explanation of theoretical concepts from related areas that are able to contribute to the investigation.

Chapter 3 describes the study's specific aims and research questions (3.1), the overall methodological approach adopted for the investigation (3.2), the research design (3.3) and preliminary steps of data processing required for the analysis (3.4).

Chapter 4 presents the results obtained in the study. This chapter is divided into three main sections: one describing results obtained in the context of a task where eye tracking and a self-report scale were exploited as data-gathering methods (4.1); one describing results based on the use of TAPs (4.2); and a third section where data obtained in these two contexts is contrasted (4.3).

Chapter 5 discusses the results obtained in the study, establishing a link with relevant theoretical concepts and empirical findings arrived at in previous work.

Chapter 6 concludes this thesis by providing final remarks on the findings obtained (6.2 - 6.4), by presenting a final discussion on the concept of cognitive effort (6.5) and on implications for practice (6.6), by describing strengths and limitations of the investigation (6.7), and by providing directions for future research (6.8).

Chapter 2. Review of Literature

A number of preliminary questions merit attention before investigating the expenditure of cognitive effort in PE. These include how PE can be operationally defined, what exactly is understood by cognitive effort and how cognitive effort can be estimated. This chapter is aimed at addressing these questions as well as at providing a review of results from relevant previous research in the area. The remainder of the chapter is organised as follows: a description of the cognitive processes involved in PE is provided in section 2.1; in section 2.2, theoretical considerations regarding the concept of cognitive effort and the ways in which this concept can be estimated are presented; methods traditionally used in previous research to assess translating difficulty and translation quality are briefly described in section 2.3; relevant results obtained in previous PE research are reviewed in section 2.4; and, finally, a summary of these results is provided in section 2.5 together with a discussion on desiderata in previous work.

2.1. The Cognitive Process of Post-Editing Machine Translation

Gouadec (2010, 26) defines PE as 'checking, proof-reading and revising translations carried out by any kind of automatic translation engine'. While this definition fits most kinds of professional settings where PE is implemented, psycholinguistic aspects of the activity and the mental operations it entails – elements directly relevant to the present thesis – are less readily definable and require empirical evidence to be fully understood.

To date the most comprehensive psycholinguistic study on PE has been conducted by Krings (2001),[1] whose research aims involved comparing the effort

[1] This is an English translation of a postdoctoral thesis submitted in 1994, based on experiments conducted in 1989-1990.

invested in traditional translation (from scratch) and in PE, and examining the feasibility of post-editing MT output with no access to the ST. Krings regards PE as a form of machine-assisted translation characterised by a distribution of attention across three textual components: the ST, the MT output and the emerging target text (TT) (ibid., 166-167). He suggests that the overall process of post-editing MT consists of constantly shifting between ST- and MT-related processes, i.e. with TT production being anchored predominantly on the MT output or predominantly on the ST, forming a triangular relationship, as illustrated in Figure 1.

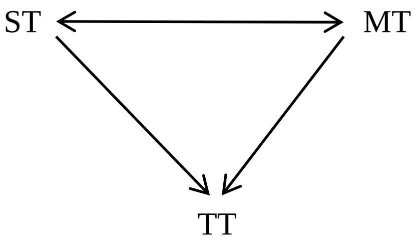


Figure 1 - Triangular relationship between ST, MT output and TT - proposed by Krings (2001, 167). Reproduced with permission from Kent State University Press

Using TAPs as an elicitation method, Krings set out to describe the different cognitive processes (i.e. mental processing steps) and sub-processes that occur during the activity. He assumed that cognitive processes in PE would be related either to one of the three textual components mentioned above or to certain elements outside this triangle, such as reference sources.

To obtain information on text comprehension processes, Krings adopted a theoretical model of reading proposed by van Dijk and Kintsch (1983). This model is referred to as ‘a textual *analysis* model’ rather than simply a model of ‘text *comprehension*’ (Krings 2001, 234). This is because this model attempts to cover analytical decisions on how to act upon the information read in the text, going beyond just comprehension. Indeed, recent research in translation (Jakobsen and Jensen 2008) indicates that different modes of reading (e.g. reading just for comprehension or reading for translation) are associated with different levels of cognitive processing, with task time and the count and duration of eye fixations on the text varying as a function of reading mode (see section 2.2.3.1 for a definition of ‘eye fixation’). By adopting van Dijk and Kintsch’s model of textual analysis, Krings seemed to take this cognitively diverse nature of reading into account.

Van Dijk and Kintsch (1983) highlight that, differently from most text comprehension frameworks, theirs is not based merely on linguistic structural ‘levels’ –

i.e. phonology, morphology, syntax, etc. –, but rather on comprehension strategies. These strategies include the formation of mental statements (propositions) and the use of schemata (i.e. knowledge previously acquired by the readers). Van Dijk and Kintsch exploited strategies as opposed to just linguistic levels because textual elements belonging to different levels are not processed in isolation. In other words, processing a word involves understanding the context surrounding the word and hence might involve understanding the clause, the sentence and so forth; a notion that might render too simplistic a model that regards e.g. morphological, lexical and syntactical processing as isolated blocks.

Building from van Dijk and Kintsch's (1983) framework, Krings (2001) proposed a set of cognitive processing components deemed to be involved in textual analysis in PE. In brief, these components are as follows:

i) word recognition;

ii) concept formation (mentally forming a concept);

iii) morphology and syntax;

iv) simple and connective proposition formation;[2]

v) constructing coherence (establishing meaningful links between different arguments in the text);

vi) text basis, inferences and elaborations (constructing a 'semantic representation of the text' (ibid., 242), which includes implicit information and concepts elaborated-on by the reader);

vii) text, macrostructure, superstructure (analysing the text's global structure); paratext (analysing any accompanying material that is not part of the main body of text, such as pictorial information);

viii) knowledge (comprehension processes relying on knowledge of the text genre, knowledge of the subject matter acquired through previously read parts of the text itself, or 'world knowledge', deemed to involve kinds of knowledge not comprised in the previous two types); and

ix) pragmatics and diasystematic markers (recognising and taking account of pragmatic aspects of the text) (ibid., 233ff.).

[2] In the formulation adopted by Krings (2001), simple propositions consist of predicate-argument relationships governed by semantic roles, e.g. agent or patient. Connective propositions would then consist of groups of simple propositions linked by conjunctions, expressing ideas of, for example, condition or time (see Krings 2001, 240).

To reveal the overall distribution of cognitive processes comprised in PE, Krings (2001) also took into account processes not directly linked to text comprehension but to the editing activity itself, such as ‘compare an element of the source text with an element of the machine translation’ (ibid., 514ff.). Krings then provided an extensive catalogue of the cognitive processes involved in PE by grouping these processes into eight categories:

i) processes related to the ST;
ii) processes related to the MT output;
iii) processes concerning TT production;
iv) TT evaluation-related processes;
v) reference-work-related processes;
vi) physical writing processes;
vii) global task-related processes; and
viii) non-task-related processes.

Krings (2001) regarded the physical implementation of edits as separate from their mental planning (coded under TT production), hence physical writing processes being treated as a separate category. Global task-related processes comprise processes that pertain to the activity as a whole, such as ‘re-orient oneself in the text’. Non-task-related processes, in turn, comprise processes that pertain to the context in which the experiments were conducted, such as ‘speak to experimenter’ (ibid., 522).

The proportion of PE processes falling into each of the processing categories proposed by Krings (2001) (corresponding to tasks carried out with access to the ST) is presented below in Table 1. As can be seen, the actual task of producing the TT accounts for nearly half of the amount of cognitive processes involved in PE (43.5%), with the TT being involved in the vast majority of processes if target evaluation (9.5%) and physical writing (15.2%) are taken into account. Of the other textual components the activity involves, consulting the MT output was associated with the second largest amount of cognitive processes (9.7%) and consulting the ST the third largest amount (8.6%).

	% of processes
target production	43.5
physical writing	15.2
MT	9.7
target evaluation	9.5
ST	8.6
reference materials	7.1
global task	5.9
non-task	0.4

Table 1 - Distribution of bilingual PE processes (Krings 2001, 314).

It is worth pointing out that while Krings (2001) catalogued the different cognitive processes that take place in PE in the context of the categories presented above, binding these categories together in a systematic way to obtain a cognitive model of the activity (i.e. a step-by-step sequence of the mental steps it involves) is a goal that to date has not been achieved in the field. Nevertheless, Krings's results provide a useful framework with empirical evidence to suggest, amongst other things, that the emerging TT is indeed the main focus of attention in PE, followed by the analysis of the MT output and of the ST, respectively.

2.2. Cognitive Effort, Cognition and Cognitive Processes

To examine different aspects of PE behaviour and, in particular, the effort that goes into the activity, it is necessary to provide an account of what is understood by cognitive effort in the present study and how this effort can be measured or estimated. This section presents an overview in this respect, including a discussion on the definition of cognitive effort (section 2.2.1), an explanation of the human memory system and its connection with effort expenditure (section 2.2.2), and the rationale for different methodologies that can be used to estimate cognitive effort (section 2.2.3). A summary covering these topics is provided in section 2.2.4.

2.2.1. *Defining Cognitive Effort*

Cognitive (or mental) effort is regarded in this thesis as the amount of mental resources individuals allocate to a task, a concept which is distinct both from the time spent post-editing and from the amount of modifications implemented in the MT output (see section 2.2.1.1, below). This section provides details of how this concept is theoretically approached in different areas, including PE itself (section 2.2.1.1) and cognitive and educational psychology (section 2.2.1.2). While terminology and the angles from which cognitive effort is regarded vary quite considerably between different fields, it is noteworthy that these areas share very similar assumptions about cognitive effort. These assumptions, which are able to contribute to the present study, are restated in section 2.2.1.3 in a summarised account of how cognitive effort is approached in this thesis.

2.2.1.1. *Cognitive, Temporal and Technical Effort in Post-Editing*

Krings (2001) suggests that the general concept of PE effort consists of a combination of three sub-components: *cognitive*, *temporal* and *technical* effort. Underpinning Krings's formulation is a comparison between PE and traditional translation, where being able to tell which of the two activities requires less effort is crucial to an effective use of PE in a commercial setting. In that respect, Krings affirms that the most direct way of measuring effort in the activity is through the time spent on the task, i.e. temporal effort, which he defines as 'the time necessary to free machine-translated text of its deficiencies' (ibid., 178-179). According to Krings, a mixture of cognitive and technical effort is what underlies temporal effort, i.e. variations in PE time. He defines cognitive effort as a concept that 'involves the type and extent of [the] cognitive processes that must be activated in order to remedy a given deficiency in a machine translation' (ibid., 179). Technical effort is then regarded as 'the effort caused by purely technical operations', such as deletions and insertions (ibid.).

Cognitive effort is seen by Krings (2001) as the central variable influencing both the number of editing operations that is implemented and the total time spent on the task, i.e. technical and temporal effort, respectively. Krings also advocates a clear distinction between cognitive and technical effort. This distinction would be due to the different nature of errors that can be found in the text – some might pose little cognitive effort and yet require a large number of editing operations, while others might be cognitively demanding and yet require few edits – effects empirically observed by Koponen (2012), for example. It is also interesting to note that Krings's catalogue of PE processes allows for non-implemented mental edits as well as processes that cannot be identified based on the post-edited product alone (e.g. if the MT output is positively evaluated and left as is) to be taken into account. It could be argued that processes of this kind constitute part of the activity and can (along with edits that are both planned and implemented) also account for the levels of cognitive effort expended by post-editors. This seems to explain the emphasis Krings places on *cognitive* effort to the detriment of the other components of overall PE effort.

At this point it is worth noting, however, that while the theoretical framework proposed by Krings (2001) is a useful tool for investigating cognitive effort in PE, a more complete understanding of cognitive effort requires further information on the psychological and neuropsychological mechanisms of this concept; aspects that Krings does not tackle directly. Information of this kind is a more direct object of investigation in other areas, such as cognitive and educational psychology. An overview of the concept of cognitive effort in the context of these areas is provided below.

2.2.1.2. *General Perspectives on Cognitive Effort and its Determining Variables*

In the general context of cognitive psychology, a plethora of definitions has been provided for the concept of cognitive effort. Perhaps the most traditional of these is the one proposed by Tyler, Hertel, McCallum and Ellis (1979, 608), who say that cognitive or mental effort is 'the amount of the available processing capacity of the limited-capacity central processor utilised in performing an information-processing task'. To arrive at this definition, Tyler et al. drew upon previous work carried out by Moray (1967), who questioned a then prevalent model of brain capacity, introducing the idea that, instead of a *channel*, the brain would function as a *processor* which works out the extent of the resources to be allocated to a task based on the nature of the task itself – the amount of the utilised resources being what Tyler et al. define as *effort*, a notion that was already present in previous work by Kahneman (1973). This definition is followed in the present study; it could be seen as complementing Krings's (2001) by making it clear that the notion of cognitive effort relates to the fact that the pool of resources available to an individual is limited (see section 2.2.2, below, for further information on the nature of this limit and on how cognitive effort is connected with the human memory system).

Arguably also helpful in shedding light on the concept of cognitive effort are previous formulations from the area of educational psychology. In this area, particularly relevant for the present study is cognitive load theory (Paas, Tuovinen, Tabbers and van Gerven 2003; Sweller, Ayres and Kalyuga 2011), a field of research that attempts to identify the cognitive demand of learning tasks as a way of proposing more efficient strategies to tackle the teaching of new information. Here, of particular relevance are questions such as how individuals acquire knowledge by means of performing a task and what the impact of this acquisition is across time, where it can be generally expected that larger volumes of acquired information lead to less effort expenditure.

According to cognitive load theory scholars, *effort* is one of the components of an overarching construct referred to as *cognitive load*, a broader concept that involves three variables: mental load, mental effort and performance (Kirschner 2002, 4). Mental load concerns the intrinsic difficulty posed by the task. Mental effort relates to the actual amount of cognitive resources dispensed by individuals when performing the task. Performance concerns the resulting interaction of mental load and effort, i.e. how participants approach the task and how successful they are in doing so, which can be measured, for example, through the amount of errors they make (Paas et al. 2003, 64).

Based on cognitive load theory, the load posed by a task can be further classified as being of two different types: *intrinsic* and *extraneous* (Sweller, Ayres and Kalyuga 2011, 57). As a theory that deals essentially with the learning process, these categories are usually explained in view of didactic materials, where intrinsic cognitive load is the load posed by the nature of the problem itself and extraneous cognitive load is the load associated with the way in which the problem is presented. When drawing a parallel between this classification and PE, the complexity/difficulty of the ST and the quality of the MT output could be seen as representing intrinsic cognitive load. Extraneous cognitive load would be posed by external factors such as the editing tool and any extra technical difficulty it might pose, for example.

Regarding the terminology employed in the context of cognitive psychology and cognitive load theory, it is worth noting that cognitive *effort* and cognitive *load* are often used interchangeably (see e.g. Paas and van Merriënboer 1994; Paas et al. 2003). It is also worth noting that the terms *capacity*, *attention*, *effort* and *resources* have all been used in the past to refer to the same notion of a commodity that underpins individuals' capability of performing a cognitive task (Wickens 1984, 67). For consistency with previous PE research, and as the demands of the task are estimated independently from effort in this thesis (see section 2.2.1.3), cognitive *effort* is the term adopted here (as opposed, for example, to capacity or attention), with *cognitive* and *mental* effort being regarded as the same concepts.

In addition to the aspects mentioned above, it is worth signalling that individual variation is central to any definition of cognitive effort. In exploring the idea that individuals' level of motivation and other personal traits may influence the amount of effort invested in a task, Yeo and Neal (2008) conducted two studies where they tested a series of hypotheses aimed at revealing the level of impact that subjective traits have on cognitive effort. Air traffic control tasks were used in both studies, where individuals' cognitive ability, level of conscientiousness, subjective ratings on the difficulty of the task, as well as the amount of practice acquired during the task itself were taken into account as possible determinants of what Yeo and Neal referred to as *subjective* cognitive effort, a construct that was estimated purely via subjects' own perceptions. Cognitive ability and conscientiousness were measured with psychological tests administered prior to the air traffic control tasks. Yeo and Neal's results demonstrated that subjects' individual characteristics have a considerable degree of influence on effort expenditure. In particular, conscientiousness and perceived task difficulty were associated with higher levels of cognitive effort, while cognitive ability and task practice were associated with lower levels of cognitive effort. This seems to support a notion that, more than estimating the 'type and extent of cognitive processes' (Krings 2001, 179) that take place in a task as an index of cognitive effort, one should also account for the subjective factors that may influence these processes.

With respect to Yeo and Neal's (2008) findings, it is worth noting that their study is based on a well-defined task where the goal to be achieved leaves little room for variation. Generally speaking, success in air traffic control means enabling aircraft to take off and land in safety and on time. PE, by contrast, should be regarded as an ill-defined task, as affirmed e.g. by Sirén and Hakkarainen (2002) with respect to translation. This is because, in PE, success is highly dependent on the purpose of each specific task – will the post-edited text be accessible to the general public or will it be used just for assimilation, by a restricted group? Even by providing post-editors with a task brief including clear instructions regarding the amount of editing and level of final quality expected, participants' performance and approach to the task would still be expected to vary in some degree as a function of their own perceptions of these instructions coupled with their own subjective traits. In view of this, it could be argued that subjects' individual characteristics are expected to be of even more importance in PE than in tasks whose goals are more objectively defined. For example, a particularly conscientious participant may expend a high degree of cognitive effort even in the context of a task where only light PE is required, while a particularly uncritical participant may overlook less serious MT errors even in a context where the final text should be indistinguishable from human translation. This, again, seems to highlight the importance of taking individual variation into account in PE.

In a recent attempt at arriving at a formula of cognitive effort that can be applied to a wider variety of activities, Longo and Barrett (2010) put forth a mathematical model that tentatively defines cognitive effort whilst comprehensively taking into account its various potentially influencing factors, such as subjects' intentions and their individual perceptions of the task. The formula they propose takes six factors into account: subjects' cognitive ability, subjects' arousal, subjects' intentions, involuntary bias from the task context, subjects' perceptions, and time pressure. By including all of these factors in a single model, Longo and Barrett postulate that cognitive effort is 'a function of time [pressure] and of the individual subjective status along with environmental properties' (ibid., 72). They regard environmental properties as relatively static characteristics that are not controlled by participants themselves, such as the intrinsic difficulty of the task (ibid.). Longo and Barrett's formula deliberately comprises generic components intended to render their model applicable to a wider range of areas where cognitive effort might need to be investigated. In the context of PE, it is reasonable to expect that post-editors might expend different levels of cognitive effort as a function of all six factors covered by their model, including previous-knowledge factors such as level of professional experience and familiarity with the genre and the editing tool; the intrinsic difficulty of the task; the urgency of the job; as well as bias from the task context (e.g. relating to noise or temperature), which in an experimental setting should ideally be kept to a minimum.

2.2.1.3. *Cognitive Effort in this Thesis*

In light of the formulations reviewed above, cognitive effort can be more specifically defined in this thesis as the amount of mental resources an individual allocates to a cognitive task in order to cope with the demands imposed by the task itself and with other factors such as time pressure and any influencing elements stemming from the task context, such as levels of light and noise. It is also understood here that individuals' expenditure of mental resources is influenced by their subjective traits, such as cognitive ability and level of expertise in the activity.

Regarding the actual variables acting as indicators of task demands and cognitive effort, textual features taken into account in previous correlational research in PE reflecting the complexity of the ST or the quality of the MT output (see section 2.3 for a review on some of these features) could be seen as representing the load posed exclusively by the task while aspects such as eye-movement behaviour and participants' perceptions can be exploited as indicators of cognitive effort, as defined above (see section 2.2.3 for a full discussion on how cognitive effort can be estimated). As for other factors that may influence cognitive-effort expenditure, the same instruction regarding time pressure was given to all participants in the present study: they were told to do the task in as little time as possible, which is in line with professional PE guidelines (see section 3.3.4). Characteristics of participants themselves were also taken into account in the analysis (see section 3.3.5) and physical conditions in the experimental room (e.g. levels of light and noise) were kept as constant as possible to prevent them from influencing the results obtained.

It should be mentioned at this point that the task demands, though traditionally regarded as intrinsic to the task (Kirschner 2002, 4), are also expected to vary as a function of individuals' traits. For example, for a group of post-editors with no specialism in any subject matter, post-editing a text with a general topic would be expected to be less demanding than post-editing a text from a specific technical area. This was taken into account in this thesis by choosing STs expected to be reasonably undemanding of subject-matter knowledge (see section 3.3.2). Cognitive effort and the demands of the task are nevertheless regarded here as separate concepts. The difference between them lies in the fact that the former can be estimated independently from the process of carrying out the task – based on characteristics of the task itself (e.g. linguistic complexity and MT quality) and on what is known about the individuals expected to carry it out – while the latter relates to the mental resources that are actually expended to cope with the task demands (Paas et al. 2003, 64).

2.2.2. *Cognitive Effort and the Human Memory System*

In addition to defining the concept of cognitive effort, understanding the mechanisms of the human mind connected with effort expenditure is also of interest to the present study, as this can shed further light not only on the concept of cognitive effort but also on practical aspects of its estimation (discussed in section 2.2.3). In view of this, this section aims at providing a brief description of the various aspects of human memory that have a connection with the expenditure of effort.

Most definitions of cognitive effort provided by previous research imply that some sort of component of the human mind is 'loaded' with information to be processed – see e.g. Tyler et al.'s (1979) definition in section 2.2.1.2. The concept of cognitive effort is thus closely related to how the human mind holds and manipulates information, tasks normally deemed to be functions of memory. Human memory is usually assumed to comprise three main systems: sensory memory, working memory and long-term memory (Baddeley 1999, 19). Sensory memory is responsible for briefly holding information coming through the senses so that this information can be processed further (see Anderson 2005, 173). It is then only to parts of the information coming through the senses that we actually devote *attention*.

The concept of attention is traditionally defined as ‘taking into possession of the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought [sic]’ (James 1890, 403-404). In other words, it is what allows our minds to, out of several possibilities, single out certain elements to be processed. Attending to information in this way is normally assumed to be amongst the functions of working memory, a system of short-term storage capable of manipulating information and of interacting with long-term memory, a separate system, of larger and more durable storage capacity. These two systems are related to a number of elements that surround the concept of cognitive effort, such as our ability to pay attention to a task, how much practice we acquire whilst carrying out the task and how this acquired practice may influence the amount of effort we expend when doing the same task on future occasions.

In the context of cognitive load theory and other theoretical frameworks in cognitive psychology, working memory is normally assumed to be the recipient of the aforementioned ‘load’ placed on our mental system (see e.g. Paas et al. 2003). It is important to note, however, that these are theorised systems that are in constant development as empirical investigation in the area advances. The account provided here on the functioning of human memory is largely based on the work of Baddeley (2007, 2009c), who has had some success in finding empirical evidence to support his theory.[3] In what follows, important concepts relating to working memory and long-term memory and the ways in which these two systems are linked to cognitive effort are described in sections 2.2.2.1, 2.2.2.2 and 2.2.2.3.

2.2.2.1. *Working Memory*

The term ‘working memory’ gained prominence after being used by Baddeley and Hitch (1974) to refer to ‘a limited capacity temporary storage system that underpins complex human thought’ (Baddeley 2007, 6-7). When Baddeley and Hitch introduced this concept, they were in fact building upon the then prevalent notion of short-term memory, described as a system associated with the ‘simple retention of small amounts of information, tested either immediately or after a short delay’ (Baddeley 2009a, 19). With regard to the differences between these two systems, Baddeley (2009a) affirms

[3] For alternative accounts on the functioning of human memory, see e.g. Cowan (1999) and Ericsson and Kintsch (1995).

that while the concept of short-term memory is associated mainly with *storage* capabilities, working memory goes beyond that, manipulating information and allowing the performance of complex tasks that involve, for example, reasoning, learning and comprehension.

The working memory system proposed by Baddeley and Hitch (1974) is a *multicomponent* model. It includes a central 'attentional controller' (Baddeley 2009b, 53), referred to as the *central executive*, accompanied by three other components: the *phonological loop*, the *visuo-spatial sketchpad* and the *episodic buffer* – the latter component being a more recent addition to the model (see Baddeley 2000). A diagram illustrating the relationship between these components is presented in Figure 2.

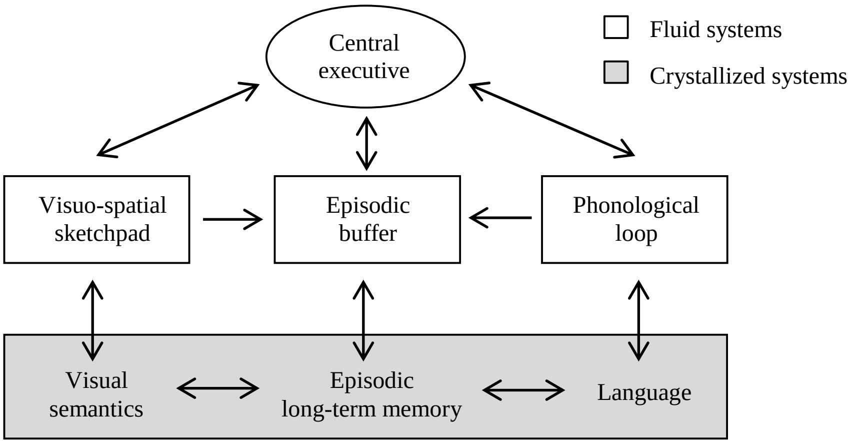


Figure 2 – Multicomponent working memory model as per Baddeley (2009c), which is a slight adaptation of Baddeley (2000). The 'fluid systems' represent working-memory components, while 'crystallized systems' refer to long-term memory. The arrows illustrate the different ways in which all these systems can interact. Reproduced with permission from Elsevier

The phonological loop is responsible for holding acoustic or speech-based information while the visuo-spatial sketchpad holds visual or 'spatially encoded' items (Baddeley 2009b, 44). The episodic buffer acts as a link between all components of working memory as well as between working memory (i.e. the 'fluid systems' in Figure 2), long-term memory (the 'crystallized systems', in Figure 2) and information coming from the senses (ibid., 57). According to Baddeley (2007, 13), the episodic buffer allows visual and verbal information to be integrated, a known capability of the human mind that was deemed to be missing in earlier versions of his working memory model (see also Cowan 2005, 37-38).

Working Memory and Error Detection in Texts

Certain components of working memory have been empirically shown to have a direct impact on (monolingual) text error detection (Larigauderie, Gaonac'h and Lacroix 1998), an activity assumed here to be relatively similar to PE (as also assumed, for example, by Temnikova 2010). It has been shown that both the central executive and the phonological loop are involved in detecting errors in texts, though with these components having slightly different functions. Larigauderie, Gaonac'h and Lacroix conducted a series of experiments where the performance of either the phonological loop or the central executive was deliberately disrupted during the task by asking participants to pronounce either a specific or a randomly chosen number at set intervals of time (which is expected to disrupt the phonological loop and the central executive, respectively). Results obtained with these experiments demonstrate that disrupting the performance of the phonological loop has a greater impact on the detection of errors above the word level, i.e. errors whose correction requires examining the link between words, such as semantic and syntactic errors. The phonological loop was not found to affect the detection of typographical or orthographic errors (i.e. letter sequences which are incorrect phonologically and errors that make phonological sense but go against spelling conventions, respectively). The importance of the phonological loop for the detection of errors that stretch across longer textual spans can be explained by the fact that this component is responsible for keeping previously decoded verbal information available for further processing – e.g. correcting a syntactic error may involve remembering the beginning of a sentence whilst also focusing on the sentence's end, which indeed seems central to PE.

Disrupting the performance of the central executive had a greater impact on task performance as a whole, with typographical errors being the only type of error whose detection rate remained unaffected in Larigauderie, Gaonac'h and Lacroix's study. The detection rate of orthographic errors was also found to be lower when the functioning of the central executive was disrupted. According to Larigauderie, Gaonac'h and Lacroix, this is because finding a solution for orthographic errors requires searching for the words' correct spelling in long-term memory, a task expected to involve the central executive.[4]

[4] These experiments were carried out prior to the inclusion of the episodic buffer in Baddeley's model of working memory. Based on a more recent formulation, the episodic buffer could also be assumed to play a role in the retrieval of this type of information (see Baddeley 2007, 139ff.).

Working Memory Capacity

Another aspect of the human memory system that is relevant to the present study is the notion that the capacity of working memory varies from person to person (Baddeley 2009b, 58). Cowan (2005, 3) defines working memory capacity as ‘the amount [of information] an individual can hold in mind at one time’. The idea that there is only so much the human mind can attend to at a time is related to early work carried out by Miller (1956), who suggested that we are only able to hold between five and nine items in the attentional focus. Since then there has been much debate on the nature of this capacity limit and its implications (see e.g. Cowan 2010). While discussions on this topic are ongoing, it is generally accepted that a capacity limit exists across a variety of tasks, and in a number of previous studies it has been found that individuals with high capacity have better performance in a range of activities, such as reading comprehension and computer programming (Daneman and Carpenter 1980; Shute 1991).

In the context of PE, given the importance of working memory for text error detection (Larigauderie, Gaonac’h and Lacroix 1998) and for cognitively demanding activities in general, there is arguably strong reason to suspect that working memory capacity may also play a role in post-editors’ performance. Some evidence in this respect has been found in the context of monolingual text revision, for example (see McCutchen 1996). In view of this, subjects taking part in the present study had their working memory capacity measured in an attempt to control for potential effects of this variable.

Recent research suggests that working memory capacity can have two underlying facets: a limit in the amount and quality (i.e. acuity or accuracy) of information that can be held in the attentional focus, referred to as *scope of attention*, and the ability to mentally select the most important items to hold in the attentional focus to the detriment of less important ones, referred to as *control of attention* (Chow and Conway 2015; Cowan et al. 2005). Both these facets are expected to play important roles in PE, for example in controlling the division of attention between ST, MT output and TT, and in holding information in the attentional focus so that errors involving long textual spans can be corrected. The test used here to measure working memory capacity (see section 3.3.4.1) is traditionally regarded as an index of control of attention, though previous research suggests it also reflects attention scope (Cowan et al. 2005).

2.2.2.2. *Long-Term Memory*

Baddeley (2009b, 10) describes long-term memory as ‘a system or systems assumed to underpin the capacity to store information over long periods of time’. It is generally accepted that long-term memory comprises two components: explicit memory and implicit memory (Squire 1992). Explicit memory is responsible for storing specific information that we tend to associate with a traditional notion of memory, such as remembering birthday dates (Baddeley 2009c). Implicit memory is related to a storage process that takes place implicitly as we learn an activity, such as learning how to ride a bicycle and being able to ride it again afterwards without being able to point out the specific pieces of information in our brain that make this possible (ibid.).

Explicit memory can be further sub-divided into two sub-systems: semantic and episodic memory (Tulving 1972). Semantic memory is responsible for storing ‘knowledge of the world’ (Baddeley 2009c, 11) (e.g. the meaning of words), sensory information (e.g. what different types of food taste like) and knowledge about how to act in specific social situations (Baddeley 2009c). Episodic memory is related to our ability to remember specific events or episodes. To illustrate the difference between episodic and semantic memory, Baddeley (ibid.) uses the example of learning of someone’s death. Once you become aware that someone has died, this piece of information is stored in semantic memory. Being able to remember the moment when you learned of the person’s death, on the other hand, is associated with the role played by episodic memory. Catling and Ling (2012, 60) highlight that storage and retrieval processes related to episodic memory require contextual information, such as place and time. Semantic memory is not associated with this type of knowledge as it does not concern specific events, but rather general information *about* things.[5]

It goes without saying that knowledge *about* things is crucial to most kinds of linguistic activities, including PE. From knowledge that allows post-editors to make lexical substitutions to knowledge regarding the subject matter of the texts being post-edited, all of this is information held in semantic memory and that is expected to be accessed during the activity. It would also be expected that, in addition to this kind of explicit knowledge, experience in the task would influence post-editors’ behaviour in some way through a sub-conscious acquisition of knowledge. As mentioned above in

[5] The correspondence between all of these sub-systems of long-term memory and the ‘crystallized systems’ illustrated by Baddeley in Figure 2 is not straightforward, but it can be assumed that, other than episodic memory (included in Baddeley’s model as a crystallized system), the other sub-systems of long-term memory mentioned above would be grouped under the label of ‘visual semantics’ or ‘language’ (see Figure 2) depending on the nature of the information stored: linguistic or visual.

the example of learning how to ride a bicycle, information of this kind is assumed to be stored in implicit memory, reflecting the notion that the more one performs an activity, the more implicit knowledge is acquired and stored, rendering this person better equipped to carry out the same activity in the future. This is connected to the idea that at certain moments of the task post-editors may simply act intuitively instead of analytically, a notion discussed below in describing the central executive and the distinction between automatic and controlled processes.

2.2.2.3. *Attention and the Central Executive*

The central executive, the component of working memory that controls the allocation of attention, is crucial to understanding the underlying mechanisms of cognitively expensive PE processes. Baddeley (1996) predicts four possible functions for the central executive: focusing attention, switching attention, dividing attention, and establishing a link between working memory and long-term memory. Of these proposed functions, the ones regarding attention are based on a framework put forth by Norman and Shallice (1986), who assume our actions can be either consciously controlled by a supervisory attentional system, or carried out automatically, based on ingrained habits.

Controlled processes ‘require attentional resources’ (Norman and Shallice 1986, 2), so they can be deemed to be directly linked to the expenditure of cognitive effort and to Baddeley’s central executive,[6] which is analogous to Norman and Shallice’s supervisory attentional system (Baddeley 2009b, 67). Automatic processes, on the other hand, do not rely on cognitive processing resources (Schneider and Shiffrin 1977a, in Norman and Shallice 1986), so it is safe to assume they do not entail cognitive effort. The question of what specific actions are supervised by the central executive, however, is a more contentious issue. Of the four functions of this component predicted by Baddeley, only one is fully supported by empirical evidence: the role of focusing attention (Baddeley 2007, 138). The roles of dividing attention and of establishing a link between working memory and long-term memory, though not proven as functions of the central executive, are described by Baddeley as promising possibilities, being the object of ongoing investigation. Little evidence, however, is available to suggest that the central executive is indeed responsible for attentional switches. Here, Baddeley highlights that switching attention from one task to another may on some occasions be ‘attentionally demanding’ and on some occasions an automatic process (Baddeley

[6] Indeed, Baddeley’s central executive is normally regarded as the component responsible for ‘allocating cognitive resources’ in translation, as pointed out by Hvelplund (2011, 44).

2009b, 56). These four potential roles of the central executive are examined below in detail in view of PE.

Attentional Focus

When analysing the role played by the central executive in translation, Hvelplund (2011, 45-46) highlights that focusing attention (i.e. in the sense of 'being focused on' rather than 'switching focus to') is necessary, for example, for ST comprehension and for TT production. He draws on work by Gile (1995) to suggest that enough time should be spent concentrating on the ST for the translator to come up with a plausible interpretation of its meaning. Similarly, TT production should also be attended to long enough to allow the translator to check if the rendition of the ST in the target language is adequate. The same in this respect could be assumed for PE, though with the MT output constituting an extra element to be focused on. Amongst other things, focusing long enough on the evaluation of the MT output is necessary to decide if a modification is needed and, if so, what this modification should be – a process which, abandoned prematurely, would be likely to incur lower levels of post-edited quality.

Attentional Division

A traditional example of attentional division is the act of driving and speaking on the phone at the same time (Baddeley 2009b, 55). In PE, attentional division would be expected to occur throughout the task, such as when both local and global aspects of the text are attended to simultaneously (e.g. the grammatical correctness of a specific phrase and how this phrase fits the general context or tone of the text). Indeed, previous work in (human) translation revision highlights that attending to both local and global aspects of the text at the same time can be particularly demanding (Mossop 2007, 38), which could be seen as a result of the limited processing capacity of the central executive.

Attentional Switching

In PE, attentional switching would be expected to take place fairly frequently – e.g. when switching between the left- and the right-hand side of the triangular relationship proposed by Krings (2001) involving the ST, the MT output and the TT (see Figure 1, p. 6). As previously mentioned, however, little evidence is available to support the involvement of the central executive in attentional switching of this kind, implying that on some occasions these switches may occur through automatic processes (Baddeley 2009b, 56).

The Link between Working Memory and Long-Term Memory

The fourth function of the central executive initially predicted by Baddeley (1996) is that of establishing a link between working memory and long-term memory. It has been shown by Brener (1940) that after being exposed to a sample of words we are able to recall a larger number of these words if they form a sentence. When these words are part of a sentence, our recall capability is of approximately fifteen, as opposed to five if the words are presented in isolation (Baddeley, Vallar and Wilson 1987). As fifteen words is beyond the capacity of the phonological loop (Baddeley 2009b, 56), this suggests that the syntactic and semantic characteristics of a sentence are able to enhance our working memory span, allowing us to recall a larger number of items immediately after being exposed to them. Since it is generally accepted that information on syntax and semantics is stored in long-term memory, this implies that working memory and long-term memory are in constant interaction, which is illustrated in Figure 2 (p. 17) by the vertical arrows linking working memory (the 'fluid systems') with visual and linguistic information stored in long-term memory (the 'crystallized systems'). The episodic buffer is the component of working memory that is in charge of *binding* (ibid., 57) these different kinds of information into episodes, i.e. memory sequences involving both visual and speech-based information. In the context of PE, an example of such an episode would be remembering how a solution for a recurrent problem was found in the past – e.g. which reference source was used, if a specific section of a website was consulted, or what definitions were looked up in the dictionary.

While Baddeley predicted the interaction between working memory and long-term memory to be overseen by the central executive (i.e. through the episodic buffer), there is evidence that this might not be the case on all occasions and that the connection between long-term memory and the three components of working memory may also occur automatically (Baddeley 2007, 314-316). This is again reflected in the lower vertical arrows in Figure 2 (p. 17), which denote that a direct connection with the central executive is not required for a link between working memory and long-term memory to be established.

Automatic and Controlled Processes in Translation and Post-Editing

The difference between automatic and controlled processes has implications for the present study on a number of levels. First, it could be hypothesised that when large ratios of automatic processes take place in PE lower levels of cognitive effort are expended. Second, controlled processes can become automatic with practice (Schneider and Shiffrin 1977b), so experienced participants could possibly count on such processes to a larger extent, being potentially more efficient in their approach to the task – i.e. by achieving better or equal results in comparison with inexperienced participants but doing so under less effort, as observed by Guerberof (2014) in terms of PE time.

Jääskeläinen and Tirkkonen-Condit (1991) looked at the automation of cognitive processes in translation by comparing experienced and inexperienced participants. They contrasted TAPs (think-aloud protocols – see section 2.2.3.3) produced by these two groups as a way of revealing the incidence of automatic processes in the activity. Jääskeläinen and Tirkkonen-Condit drew on the work of Ericsson and Simon (1984), who suggest that only controlled processes are available in working memory for verbalisation. Based on this rationale, when a process was verbalised by inexperienced participants but omitted by experienced ones, this process was regarded as automatic. Jääskeläinen and Tirkkonen-Condit suggest that not only are certain translation processes inherently automatic for more experienced translators, but also that some of their processes can become automatic during the task. Automation that took place during the task was observed, for example, when participants dealt with aspects relating to the text genre, such as register and style. Jääskeläinen and Tirkkonen-Condit observed that these aspects required both experienced and inexperienced participants to make critical translation decisions, but that inexperienced participants tended to deal with these issues on the fly, as they came across specific problems. Experienced participants, on the other hand, tended to spend some time considering aspects of this kind at the beginning, with this initial consideration allowing these issues to be dealt with automatically in the remainder of the task.

Related to the difference between automatic and controlled processes is the notion of *proceduralisation*, defined as ‘the process of converting the deliberate use of declarative [i.e. explicit] knowledge into pattern-driven recognition’ (Anderson 2005, 289). Declarative knowledge is the type of knowledge stored in explicit memory, i.e. information we are aware of knowing (ibid., 240-241). Pattern-driven recognition refers to the capability of, instead of planning one’s actions and thinking them through, simply recognising what is the right thing to do given the context (ibid., 289). It follows that automatic processes are not necessarily linked to information stored sub-consciously in implicit memory. Through proceduralisation, post-editors may turn what initially was a conscious, thought-through process into something procedural, which seems related to the phenomenon observed by Jääskeläinen and Tirkkonen-Condit (1991).

In brief, based on current theorisations of the human memory system, cognitive effort is related mainly to working memory and its attentional controller, the central executive. Only tasks that require controlled attention (i.e. tasks that are not automatic) are expected to rely on the limited processing capacity of this controller, so it can be assumed that cognitive effort is associated predominantly with non-automatic processes overseen by the central executive. At this point it is worth noting, however, that specific information regarding which sub-tasks of PE tend to be carried out automatically and which tend to be controlled is currently unknown to the field. Because of this, other than assuming that the expenditure of cognitive effort entails controlled attention, the question of where automation specifically takes place in PE cannot be directly addressed here. Nevertheless, the notion that PE, as most cognitive activities, is expected to involve some degree of automation is an important one to bear in mind when analysing empirical results on behavioural aspects of the activity.

2.2.3. *Methods for Investigating Cognitive Processes*

So far the concept of cognitive effort has been described and an explanation of the human memory system has been presented as the psychological mechanism that underlies this concept. However, an important aspect of cognitive effort still needs to be discussed: how, in practical terms, this concept can be investigated.

Setting out to estimate cognitive effort poses a number of challenges since not only is it a concept that cannot be measured directly, but it is also largely determined by individual characteristics (see section 2.2.1). Substantial research on methodological aspects of this kind has been carried out in the field of workload, a term normally used to refer to a concept similar to what is regarded here as cognitive effort.[7]

It is generally accepted that workload can be estimated in one of three ways: subjectively, psychophysiologically or through performance-based measures (Cain

[7] Cain (2007, 3) defines workload 'as a mental construct that reflects the mental strain resulting from performing a task under specific environmental and operational conditions, coupled with the capability of the operator [in this thesis, research participants] to respond to those demands'. Even though the term workload is normally used in the context of a different type of task (e.g. operating computer systems) the definition of this concept is very similar to what is understood here by cognitive effort, so the workload literature is also taken into account in this thesis, especially with regard to measurement methodology.

2007, 6). In the subjective group are, for example, strategies based on rating scales. Psychophysiological strategies comprise methods such as eye tracking, electrocardiography (i.e. equipment that produces heart-rate measures) and electroencephalography (i.e. measures reflecting brain activity on the scalp). Performance-based strategies comprise methods such as primary- and secondary-task experiments where the capability of performing concurrent tasks is measured, as in the experiments carried out by Larigauderie, Gaonac'h and Lacroix (1998) (see section 2.2.2.1).

In early research on workload measurement methodology, O'Donnell and Eggemeier (1986) proposed a set of conditions that should ideally be met to guarantee the usability and validity of different measures. These conditions involve aspects such as sensitivity, diagnostic power, subject acceptance and implementation requirements (ibid., 2-5). Cain (2007) extended this list with a number of conditions, including the method being timely and capable of capturing transient changes in workload, being replicable, and being insensitive to procedural artefacts (ibid., 4). Based on these conditions, psychophysiological measures would have the advantages of avoiding potential subjective distortion and, in most cases, having little interference with the task (see May, Kennedy, Williams, Dunlap and Brannan 1990, 77). On the other hand, it could be argued that psychophysiological measures are relatively poor for the purpose of gaining insight into qualitative details of the mental processes taking place. For this reason and in view of the general complexity involved in the estimation of cognitive effort, adopting a mixture of methodologies seemed desirable in the context of this thesis.

Cognitive effort is estimated here based on a combination of eye-tracking measures and subjective ratings, with TAPs being exploited as a way of obtaining a more detailed account of cognitive processes and of what goes on in participants' minds. An examination of the theoretical grounding of these three data types is presented below, with section 2.2.3.1 focusing on eye movements, section 2.2.3.2 focusing on subjective ratings and section 2.2.3.3 focusing on TAPs.

2.2.3.1. *Eye movements*

The use of eye tracking to investigate cognitive processes is well established in reading research (Rayner 1998) and it has also gained popularity in research on translation (see e.g. Alves, Pagano and da Silva 2010; Jensen, Sjørup and Balling 2010; Hvelplund 2011) and on PE (Carl, Dragsted, Elming, Hardt and Jakobsen 2011; O'Brien 2011).

The theoretical rationale for this method is based on two assumptions from the field of cognitive psychology: the immediacy and eye-mind assumptions (Just and Carpenter 1980). The immediacy assumption posits that ‘the interpretations [of a text] at all levels of processing are not deferred; they occur as soon as possible’ (ibid., 330). The eye-mind assumption postulates that ‘there is no appreciable lag between what is being fixated [by the eyes] and what is being processed’ (ibid., 331). According to these two assumptions, by fixating the eyes on a given textual element during reading, the reader would necessarily be engaging in the mental processing of the element being read, which should happen with no processing deferral, i.e. with the physical activity of reading and the mental processing of the text taking place at the same time.

The use of eye-tracking data as an index of cognitive effort normally involves metrics reflecting two types of eye-movement behaviour: *fixations* and *saccades*. Fixations are defined as ‘eye movements that stabilize the retina over a stationary object of interest’ (Duchowski 2007, 46). According to a bottom-up model of visual attention (i.e. a model that assumes that the element looked at is the main parameter or starting point for processing) the eyes are deemed to first scan the whole area of a stimulus in low resolution before focusing on a specific element within this area (ibid., 5). The area considered to be the target of attention at a given point in time, seen in high resolution, is referred to as the foveal area, while the area immediately around it, seen in lower resolution, is referred to as the parafoveal area (ibid.). The region immediately around the parafoveal area is what is known as peripheral vision (ibid.). Fixations enable a given area of interest within a stimulus to be held in focus (in foveal vision), whilst the area around it is kept in low resolution. When investigating human cognitive activity, fixations are normally taken into account in terms of their duration and count within a given area of interest in a stimulus, where the longer and the more frequent they are the larger is the volume of cognitive processing they are assumed to represent (Radach, Kennedy and Rayner 2004; Alves, Pagano and da Silva 2010; O’Brien 2011), allowing them to be used as an indicator of cognitive effort.

Saccades are defined as the ‘eye movements used in repositioning the fovea [central part of the eye] to a new location in the visual environment’ (Duchowski 2007, 42). While fixations enable an object of interest to be seen in high resolution, saccades enable shifts from one object of interest to another, moving foveal vision across different areas of a stimulus. Rayner cites a number of studies that show that our eyes are unable to capture visual information during saccadic movements, a phenomenon referred to as *saccadic suppression* (1998, 373). This happens because saccades take place in too short an interval of time – ranging between 10 and 100 milliseconds (Duchowski 2007, 42) – to allow for any visual information to be gathered. There is evidence to suggest, however, that the processing difficulty of a text influences both fixations and saccades (Rayner 1998, 376).

The eye-mind and immediacy assumptions, described above, form part of a *processing* model of eye-movement control, which regards the stimulus and its mental processing as the main factors controlling how and when we move our eyes. Naturally, the use of eye tracking to investigate cognitive processes is based on this type of model, but the eye-mind and immediacy assumptions have been a target of criticism in previous research. In this respect, Hvelplund (2011, 68) highlights the concepts of overt and covert attention put forth by Posner (1980, 5-6). Based on these two concepts, not only is the mind able to process what is in foveal vision, but also to *covertly* process elements that are not held in the visual focus. This contradicts the assumption that vision and mind are necessarily linked, as suggested by Just and Carpenter (1980), since according to this opposing view we would be able to mentally process elements that we do not directly look at when reading. Further criticism of the eye-mind and immediacy assumptions is based on studies carried out by Irwin (1998) and Inhoff and Radach (1998), who show that the lexical processing that takes place during a fixation continues during a subsequent saccade. Here, the fact that mental processing takes place during saccades, when no visual input is transmitted to the brain, also goes against the hypothesis that vision is *necessarily* linked to cognition, with no processing deferral.

In response to these criticisms, Rayner (1998, 389) affirms that ‘evidence is more consistent’ for models that assume a direct link between eye-movement behaviour and information processing in reading. In the context of translation, Hvelplund (2011, 69) acknowledges that one cannot be entirely certain that when a research participant looks at a given word in a text this word is being cognitively processed. However, he highlights that any potentially interfering effects in this respect should be minimal and that their size would not be expected to vary between subjects, consequently not having a harmful impact on results. Based on this same reasoning, on the established tradition of eye tracking in previous research and on the little interference this method has with the task, eye-movement data is regarded here as suitable for an investigation of cognitive effort in PE, constituting part of the mixed-method approach adopted for the study.

2.2.3.2. *Subjective Ratings*

One of the simplest ways of estimating the amount of cognitive effort expended by a group of participants is to ask them to report it. This is usually done with the use of rating scales, a strategy largely implemented in the context of workload measurement and educational psychology, and also employed in the present study. The theoretical grounding of this method is based on the fact that differences in task demand are expected to produce a feeling of effort that subjects are able to report (O'Donnell and Eggemeier 1986, 7). In the context of workload measurement, a number of rating scales have been designed for this purpose. Perhaps the most traditional of these are multidimensional, i.e. scales that require participants to provide ratings on a number of different components. This is the case, for example, of NASA-TLX (National Aeronautics and Space Administration - Task Load Index) (Hart and Staveland 1988) and SWAT (Subjective Workload Assessment Technique) (Reid and Nygren 1988).

NASA-TLX has been recently used by Sun and Shreve (2014) as an index of task difficulty in translation. The original version of this index has six dimensions: mental demand (complexity of the task), physical demand (the amount of physical activity required by the task), temporal demand (how much time pressure is felt as a result of the task pace – e.g. when operating a computer system), performance (how satisfied subjects feel with the results obtained), effort (how hard subjects have to work, mentally and physically) and frustration (how 'insecure, discouraged, irritated, stressed and annoyed' subjects feel vs. 'secure, gratified, content, relaxed and complacent') (Hart and Staveland 1988, 169). Sun and Shreve used only four of these dimensions, as physical and temporal demands were not applicable in the context of their study. They observed high levels of consistency in the scores provided for each of these dimensions and found that the overall sum of these scores can be used as a reliable indicator of task difficulty in translation.

However, having to assess a number of different dimensions could be seen as a practical disadvantage of the scales described above in terms of both time costs and analysis complexity. As an alternative in this respect, educational psychologists and cognitive load scholars have developed unidimensional scales that have also been proven to be reliable and sensitive to variations in task demands (see Sweller, Ayres and Kalyuga 2011, 73). This seemed like a desirable approach to exploit in the present study. The scale used here has been developed by Paas (1992). It asks subjects to provide ratings on ‘mental effort’, which is regarded in this thesis as synonymous with cognitive effort (see section 3.4.3 for details of how this scale was implemented).

Criticisms received by rating scales normally pertain to the subjective nature of this method. Gopher and Donchin (1986, 2) affirm, for example, that ‘an operator [in this thesis, a research participant] is often an unreliable and invalid measurement instrument’. O’Donnell and Eggemeier (1986) also highlight that subjective ratings may lack diagnostic power, providing only a global account of workload, as opposed to reflecting more specific variations of this construct throughout the task. In the context of a PE study, O’Brien (2011, 201) also mentions that fatigue and boredom may render subjective ratings less reliable.

The time interval between carrying out the task and the rating event is another aspect highlighted as a potential problem of subjective ratings. It is generally assumed in this respect that the longer the interval the higher the chance of distortion in the scores provided is, because information relevant to the assessment might no longer be available in working memory (Gopher and Donchin 1986). In addition, Gopher and Donchin point out that if one intends to use estimates of workload as an indicator of task performance (i.e. the levels of quality of the results obtained), rating scales are more suitable for tasks that are not expected to rely on a large amount of automatic processes, as these processes cannot be consciously accessed by participants and therefore cannot be taken into account in their assessment.

In the context of this thesis, the time lag between the task and the rating event is not regarded as problematic, as subjective ratings in the present study were provided by participants immediately after they edited each MT sentence used in the investigation, with no delay (see section 3.3.4). The occurrence of automatic processes, which impedes the use of subjective ratings as a performance indicator, did not seem a problem here either as estimates of cognitive effort are not assumed in the study a priori as indicators of post-edited quality. In addition, it is noteworthy that the subjectivity of self-reported measures is turned into an advantage by a number of investigators. Jex (1988, 14), for example, considers subjective ratings ‘the fundamental measure, against which all objective measures must be calibrated’. The high status Jex gives to subjective ratings is explained simply by the fact that there is no single objective parameter to measure workload, potentially making subjective ratings the most direct of all measures available. This was followed by Sun and Shreve (2014) in their adoption of a subjective score to estimate task difficulty in translation and is also taken into account in this thesis by making use of subjective scales as one of the measurement strategies exploited in the investigation. Ultimately, it is worth highlighting that these estimates are triangulated in the study against other data sources.

2.2.3.3. *Think-Aloud Protocols*

While eye tracking is able to provide a relatively objective parameter for estimating the amount of cognitive resources a subject allocates to a task and while subjective ratings are perhaps the most direct estimates of the levels of effort felt by subjects, these measures arguably fail to reflect the nature of the mental connections that take place or the thoughts that go on in participants' minds. TAPs (think-aloud protocols) were frequently adopted in the past as a means of gaining access to information of this kind. This method consists of asking research subjects to describe what they are doing and what is going on in their minds when carrying out a task, which can be done concurrently with the task or retrospectively. In translation studies, TAPs have been used mainly in the context of qualitative investigations of translation processes (see Jääskeläinen 2002 for an overview). To the present author's knowledge, Krings (2001) has conducted the only TAP-based study in PE to date where the amount of cognitive processes elicited from participants' verbalisations was used as an indicator of cognitive effort.

The rationale for the use of TAPs in linguistic research is based on an assumption put forth by Ericsson and Simon (1980), who suggest that, when carrying out a task and 'thinking aloud' at the same time, 'producing verbal reports of information directly available in propositional form does not change the course and structure of the cognitive processes' (ibid., 235). Based on this assumption, TAPs would only interfere with cognitive processes when the information reported is not originally in verbal form, which would require this information to be recoded. It follows that producing verbal protocols whilst post-editing or translating would supposedly not interfere with the task in a harmful way since these are verbal activities by their own nature.

It is noteworthy, however, that the use of verbal data in translation studies research has been a target of much debate of late, with TAPs coming under criticism due to theoretical weaknesses and, to a lesser extent, as a result of empirical observation. Krings (2001) mentions three potential weaknesses that can be associated with TAPs when they are used to investigate human cognition: (i) a disconnect between verbal reports and actual cognitive processes; (ii) interference of the think-aloud condition with the main task; and (iii) the fact that TAPs may not represent a complete account of the cognitive processes that take place (Krings 2001, 220; see also Li 2004; O'Brien 2005; Sun 2011).

According to Krings (2001), a disconnect between TAPs and cognitive processes can occur as a result of three situations: when participants are asked to report processes that are no longer available in working memory (i.e. in the case of retrospective TAPs), when they are asked to report processes that are available in working memory but that are non-verbal in nature and when they are asked to report processes that would normally take place automatically (i.e. that are not available in working memory to begin with). Krings concludes that out of all types of verbal-data elicitation, concurrent TAPs would be the one where these situations would be least problematic. This is because in concurrent TAPs information is available in working memory at the time of verbalisation, which mitigates the risk of any gaps in long-term memory being retrospectively 'filled' with processes that may not have actually happened, a known danger of retrospective verbalisations (Krings 2001, 223). Krings also highlights that participants are not expected to verbalise automatic processes or non-verbal information in a translation studies context, affirming that TAPs are therefore suitable for research in this area. Here, however, he also mentions that even in the context of a task which is verbal in nature, such as PE or translation, *some* non-verbal information would be expected to be mentally accessed. Similarly, it could also be hypothesised that it is possible for participants to report or interpret automatic processes immediately after realising these processes have taken place (see Hogarth 2012, 69-70), which would question the idea that TAPs are a direct, 'raw' account of what goes on in subjects' minds.

As for the second weakness of TAPs mentioned by Krings (2001), i.e. that they may interfere with the main task, empirical findings from previous research indicate that the think-aloud condition is associated with an increase in task time (i.e. a slow-down effect) and also that the text and any problems in it are broken down into smaller pieces when participants are asked to think aloud (Krings 2001; Jakobsen 2003). Jakobsen (ibid.) also observed that texts translated under the think-aloud condition tended to be of higher quality than texts produced under normal circumstances. This finding seems to confirm a hypothesis put forth by Krings that by thinking aloud subjects become more aware of their own cognitive processes, being better equipped to reflect and act upon their reflections. Krings suggests, however, that none of these interferences would be expected to qualitatively distort the data or substantially change the structure of cognitive processes (2001, 229), though little evidence is available to date to confirm this.

Regarding the third weakness of TAPs, i.e. that they may not offer a complete account of the non-automatized mental processes that take place, Krings (2001) mentions two situations that could account for this problem: differences between participants in their willingness or ability to think aloud and the fact that various thoughts or processes can occupy working memory at the same time, which requires that only one of these processes be selected for verbalisation. While Krings acknowledges that the level of severity of these problems has not to date been sufficiently examined, he argues that TAPs will in any circumstance be richer than data produced with other methods currently available, which are not themselves expected to provide complete accounts of mental processes.

Because of the criticisms received by the think-aloud method, TAPs have experienced a decrease in popularity in the field of translation studies of late. It is worth noting, however, that few empirical results are available to suggest that the interference posed by the think-aloud condition invalidates TAP-based findings. Indeed, based on a review of the literature and on a survey carried out with translation scholars, Sun (2011, 946) concludes that ‘there is, to date, no strong evidence suggesting that TAP significantly changes or influences the translation process’. Jakobsen (2003) highlights that, even though the theoretical assumptions that support the use of the think-aloud method may need to be revisited, certain research questions may be better answered through TAPs, which can be used in combination with automatic computer-based logs of the translation process (ibid., 31).

It seems therefore that the *severity* of the problems mentioned above is the key factor in determining if the weaknesses of the think-aloud method outweigh the richness of the data it produces. Some of the aims of the present study lend themselves more suitably to the use of TAPs – such as gaining insight into the different linguistic aspects of the task post-editors attend to (see section 3.1) – so this data source is also used here. However, due to a lack of specific information regarding the size of the interference posed by this method, it seemed desirable not to assume TAPs a priori as a measure of cognitive effort. Rather, in view of the relevance of this issue for further research in translation and PE, it seemed interesting to compare TAP-based data with other data types that can be used as indicators of cognitive effort – in the present study, eye movements and subjective ratings. This has the potential to shed light on how TAPs compare to other less invasive methodologies, which constitutes one of the aims of this thesis (see section 3.1).

2.2.4. *Summary on the Concept of Cognitive Effort*

As previously mentioned, the concept of cognitive effort is understood here as the amount of mental resources individuals expend when carrying out a cognitive task. Based on the theoretical overview provided in the previous sections, a number of conclusions can be made in regard to this concept:

i) it depends on the limited capacity of individuals' mental processor (broadly speaking, working memory);

ii) it varies as a function of the task demands, participants' individual characteristics, as well as characteristics of the context, such as time pressure;

iii) it cannot be measured directly, posing a number of methodological issues for its estimation.

These conclusions are naturally of direct relevance to the present study in that they determine the best conditions in which cognitive effort can be investigated. The fact that effort expenditure depends on individuals' cognitive ability seems to count as a justification for the measurement of working-memory capacity in the study, as mentioned in section 2.2.2.1 (see also section 3.3.4.1). The fact that cognitive effort varies as a function of other individual characteristics justifies other choices made here, such as taking participants' source-language knowledge into account in the analysis (see section 3.3.4.2 for further details).

The concept of long-term memory has also been reviewed. Long-term memory is the human memory system that is in charge of storing ingrained habits that allow us to carry out certain tasks without conscious control, i.e. automatically. This system is also linked to the expenditure of cognitive effort in that it can be assumed that the more one relies directly on long-term memory, without the use of the central executive (the component of working memory that controls the allocation of attention), the less cognitive effort is expended. This happens, for example, due to expertise, which functions as a rationale for taking participants' level of professional experience into account in the analysis (see section 3.3.5).

Regarding the rationale for different methodologies that can be used to investigate cognitive effort, the overview provided above touches upon a number of methodological issues that stem from the fact that cognitive effort cannot be measured directly. This requires that some other aspect of the task be taken into account as a reflection or estimation of this construct. A discussion of three types of data that can be used for this purpose, namely eye movements, subjective ratings and TAPs, was provided. It is clear with respect to these data types that all of them have weaknesses. The first problem mentioned by Krings (2001) in regard to the validity of TAPs is that there may be a disconnect between verbal data and the actual mental processes that take place. This problem also applies to eye movements, for example, where some evidence is available to suggest that gaze behaviour and cognitive processing may on some occasions not coincide (see Rayner 1998 for an overview). Subjective ratings are arguably the most direct estimate out of the three data types considered (see Sweller, Ayres and Kalyuga 2011, 73), though this method too has been a target of criticism due to its very subjective nature (Gopher and Donchin 1986) and to potential distortions caused by boredom and fatigue (O'Brien 2011).

In view of the inherent weaknesses of different strategies, a mixture of methodologies was adopted in the present study to estimate cognitive effort. Eye-tracking metrics and subjective ratings, data types with a long tradition of research in reading and cognitive psychology, were adopted as the primary measures of cognitive effort in the study. TAPs, on the other hand, have been a more direct target of criticism in previous years, which serves as basis for one of the research aims of the study, i.e. to check to see how TAPs correlate with other data types that can be used to estimate cognitive effort (see section 3.1).

2.3. Methods of Text Analysis and Translation Evaluation

Previous research on PE effort frequently contrasts, on the one hand, some behavioural aspect of the activity (e.g. task time) with, on the other hand, characteristics deemed to indicate the complexity or translating difficulty of the ST, or the quality of the MT output. Results in this respect obtained in previous research are reviewed in section 2.4, while methods of text analysis and translation evaluation frequently used in this context are described below in sections 2.3.1 and 2.3.2.

2.3.1. *Measures of Translating Difficulty*

A number of textual indices have been used in previous research as potential predictors of translating difficulty, such as lexical frequency, indices of readability – i.e. 'what makes some texts easier to read than others' (DuBay 2004, 3) – the incidence of different part-of-speech categories in the text, and amount of non-literal expressions. Of these, indices of readability are perhaps the measures most frequently used to date as potential indicators of translating difficulty (e.g. Hvelplund 2011; Sun and Shreve 2014).

Most formulae of readability are based on word and sentence length. Flesch Reading Ease (Flesch 1948), for example, one of the most traditional of such scores, is based mainly on number of syllables and sentences per 100-word text sample (DuBay 2004, 21). Naturally, factors of this kind would be expected to reflect just superficial characteristics of the text, failing to indicate aspects such as lexical complexity or frequency. As a likely effect of these weaknesses, a history of mixed results is associated with these formulae, with a number of previous studies finding them to be unreliable (see e.g. DuBay 2004). Recent research, however, has observed that Flesch Reading Ease (Flesch 1948) is indeed linked to the perceived translating difficulty of texts, though with this effect being supported by only a small correlation – Kendall tau-b (609): − 0.141, p < 0.001[8] (Sun and Shreve 2014, 112).

The deficiencies of traditional readability formulae have motivated more sophisticated initiatives in assessing readability as well as in predicting the translating difficulty of different STs. In addition to readability formulae, Hvelplund (2011) uses word frequency (based on text corpora) and the amount of non-literal expressions in the text. Improved versions of readability formulae have also been proposed in recent research, such as the indices yielded by the Coh-Metrix tool[9] (Graesser, McNamara, Louwerse and Cai 2004), a program capable of analysing texts with respect to characteristics that go beyond the simple properties taken into account by traditional readability scores. Information provided by this tool includes polysemy (i.e. the amount of different meanings a word has), hypernymy (i.e. how general a word is) as well as a number of psycholinguistic indices, such as imagability (how easy it is to form a mental image of a word) and age of acquisition (the age from which English-speaking individuals would be expected to know the meaning of a word).

The Coh-Metrix tool plays an important role in the present study in that the indices it produces are tested here as potential predictors of cognitive effort. Potential effects of measures such as word frequency and the incidence of different part-of-speech categories are also examined; these measures are used in this thesis in

[8] Six hundred and nine (609) is the number of degrees of freedom in the analysis, followed by the correlation coefficient (- 0.141) and the p-value. In the Flesch Reading Score scale, high values indicate high readability, which explains the negative correlation coefficient with translating difficulty observed by Sun and Shreve (2014).

[9] See http://cohmetrix.com (accessed 15 February 2015).

two ways: at the text level, as a way of selecting different STs to be used in the investigation, and at the sentence level, as potential predictors of cognitive effort in PE. Readability formulae (such as Flesch Reading Ease) cannot be applied to individual sentences, so measures of this kind are used here just for the purpose of ST selection (see details in section 3.3.2). Individual Coh-Metrix indices, such as polysemy and imagability, were tested at the sentence level, based on the MT output. Further information on how sentence-level measures were obtained in the present study and which specific ones were used can be found in section 3.4.1.

2.3.2. *Methods of Translation Quality Assessment*

The expenditure of cognitive effort in PE is examined in this thesis relative to the quality of both the MT output and the post-edited text. This required the use of assessment methods that produced quantifiable results and that could also be applied to both the raw MT output and the post-edited product, so that any levels of improvement achieved through PE could be examined.

Between MT output and post-edited text, it could be argued that the MT output predetermines quality-assessment methods to a greater extent by requiring that a wider range of quality levels be taken into account – assessment scales to be used only with human post-edited output would be unlikely to account for incomprehensible text, for example, which on some occasions is possible in the case of raw MT. In view of this, a review is provided below of methods traditionally used for MT evaluation, but which can also be applied to the post-edited text. Both automatic and human methods of assessment are described.

2.3.2.1. *Automatic Evaluation Metrics*

Evaluating MT is usually costly and time-consuming, so a number of automatic methods have been proposed in the literature for this purpose in previous years. Amongst these methods are so-called automatic evaluation metrics, which normally function by measuring the degree of similarity between the raw MT output (normally referred to as the *candidate* or *hypothesis* translation) and one or more human translations used as reference. These metrics require the existence of a human reference translation in any assessing event, which inevitably limits their use from the perspective of post-editors, as PE would not be required if high-quality translations of the ST were already available. Nevertheless, methods of this kind are useful, for example, when developing MT systems, as a way of checking to see how well the system is performing based on a test dataset containing reference translations. Some of these metrics are also used after PE has taken place to measure the amount of changes implemented in the MT output, by comparing the similarity between the raw MT output and the corresponding post-edited version.

It is worth noting that even though automatic metrics of the kind described above are frequently used as measures of translation quality (e.g. Green, Chuang, Heer and Manning 2014) this assumption will depend on the quality of the reference(s), as low-quality reference translations would of course be a bad parameter for comparison (see O'Brien, Choudhury, van der Meer and Monasterio 2011). Nevertheless, provided high-quality references are used, a number of such metrics have presented strong correlations with human assessment methods in previous research (see e.g. Machácek and Bojar 2013). One of these metrics, Meteor, was used as basis for the selection of the MT output in the present study, also being tested at a sentence level as a potential predictor of cognitive effort. This has influenced the choice of STs to be used in the investigation, since only texts with pre-existing human translations could be considered for selection (see sections 3.3.2 and 3.3.3 for further details).

The automatic evaluation metrics most frequently used in previous research are briefly described below.

BLEU

BLEU (BiLingual Evaluation Understudy) (Papineni, Roukos, Ward and Zhu 2002) is one of the most traditional automatic evaluation metrics. BLEU varies between 0 (complete disparity between MT and reference) and 1 (perfect match). This metric works by examining what proportion of the MT output matches the reference, but it does not take incomplete translations into account (i.e. passages in the reference that might be missing in the MT output). BLEU has been criticised for a number of reasons. Callison-Burch and Osborne (2006) call attention to the fact that BLEU is too permissive in what it regards as a match. This metric tries to allow for mere differences in word order (between the MT output and the reference) that would not have an impact on quality, but according to Callison-Burch and Osborne it goes too far in doing that, giving good scores to sentences containing incorrect syntax. Another criticism of BLEU is that, even though it has been shown to correlate well with human judgments at the corpus level (Papineni et al. 2002), this metric is known to be unreliable for evaluating individual sentences, which has motivated a number of attempts at adapting the score for sentence-level evaluation (see e.g. Lin and Och 2004; Song, Cohn and Specia 2013).

BLEU is also known to overestimate the quality of short sentences, which stems from the fact that the score fails to take into account any additional content in the reference translation which is not covered by the MT output, as mentioned above, a situation where higher levels of match are more easily obtained for shorter MT sentences. The score applies a penalisation to short sentences as a way of offsetting this effect (Papineni et al. 2002), but potential artefacts of this bias have nevertheless been observed in previous research (see Green et al. 2014).

TER

TER (Translation Edit Rate) (Snover, Dorr, Schwartz, Micciulla and Makhoul 2006) measures the minimum number of editing operations – insertions, deletions, substitutions (of single words) and shifts (of any number of words) – estimated to be necessary to turn a machine translation into a reference translation. Since a larger number of editing operations is expected to be associated with poorer MT output, high TER scores are expected to be associated with lower MT quality. It has been previously shown that TER correlates with human judgments better than or as well as BLEU (ibid.) and that it also correlates well with processing speed in PE (number of ST words processed per unit of time) (O'Brien 2011). TER is also frequently calculated with post-edited versions of the raw MT output acting as reference translations, a variation of the metric referred to as HTER (Human-Targeted Translation Edit Rate). HTER is more frequently used in research contexts, as an indicator of PE effort (as opposed to MT quality) (e.g. Koponen 2012), i.e. a way of measuring the amount of changes implemented in the raw MT output during the PE process. It is worth noting that HTER and TER do not take actual keyboard use into account, but rather the estimated minimum number of edits (as defined above) necessary to turn the MT output into the closest reference translation (as measured by the score itself) available. A more recent version of TER, TER-Plus (Snover, Madnani, Dorr and Schwartz 2009), is also capable of taking semantic information (e.g. synonyms) into account when computing matches between the MT output and the reference.

GTM

GTM (General Text Matcher) (Turian, Shen and Melamed 2003) assesses the level of similarity between the MT output and the reference translation by means of calculating traditional measures of Precision, Recall and F-measure. Precision consists of how much of the machine-produced sentence matches the reference. Recall represents the proportion of the reference that also matches the MT output, accounting for potentially incomplete translations, i.e. the fact that even when 100% of the information conveyed by the MT output is precise (matches the reference), there might be passages of the reference translation that are not found in the MT output. The F-measure is then the harmonic mean between Precision and Recall. In previous research, GTM has been shown to correlate well with human judgments on translation quality (ibid.), with PE time (Tatsumi 2009) and with eye-tracking measures collected during PE tasks (O'Brien 2011).

Meteor

Meteor (Metric for Evaluation of Translation with Explicit Ordering) (Banerjee and Lavie 2005) is a relatively recent metric that has been developed with the goal of addressing some of the weaknesses of BLEU. This score compares a candidate translation with one or more reference translations on a word-by-word basis. It assesses the match level between the MT output and the reference translation based on both surface and stemmed word forms, i.e. accounting for the fact that, for example, 'tell' and 'telling' (a stemmed form of tell) may be equivalent alternatives depending on how the sentence is structured. In addition, Meteor takes lexical meaning into account by matching synonyms based on WordNet[10] semantic data. This metric also inspects how well ordered matched words in the MT output are in relation to the reference translation. Meteor has presented better correlations with human judgments in comparison with BLEU both at the system and sentence level (ibid.). More recent versions of the score (Denkowski and Lavie 2010, 2011) are also able to account for paraphrases and differences in punctuation (between candidate and reference translation) with no impact on meaning (i.e. differences that do not have any bearing on the assessment).

As mentioned above, Meteor is used in the present study both in the process of selecting different machine translations for the investigation (see section 3.3.3) and as a potential predictor of cognitive effort in PE (see section 4.1.2). Meteor seemed like a good metric to use here for these two purposes as it is able to take semantic information into account in the assessment. Even though TER-Plus also makes use of semantic information when contrasting the reference translation and the hypothesis, Meteor

[10] WordNet is a freely available lexical database of English – see http://wordnet.princeton.edu/ (accessed 19 February 2012).

seemed like a better choice in this respect as more recent versions of the metric are available.

2.3.2.2. *Machine Translation Quality Estimation*

A currently emerging area in the field of MT evaluation is quality estimation. Differently from the metrics described above in section 2.3.2.1, quality estimation systems do not require pre-existing reference translations to assess the MT output. These systems are developed by monitoring the PE process of a given number of sentences and keeping track of textual characteristics (of the ST and the MT output) that are associated with high levels of PE effort – in terms of time spent post-editing or HTER, for example. This information can then be exploited for predicting the effort posed by new pairs of ST and corresponding MT. In other words, once a given amount of sentences is post-edited (i.e. in the process of developing the quality estimation system), the system will then use textual characteristics that correlate with PE effort as a basis to predict the quality of unseen MT outputs (and corresponding STs) for which no reference translations are available and which have not yet been post-edited. This assessment strategy is closely related to the present study in that it relies on textual predictors of PE effort. Blatz et al. (2004) show the value of using predictors based both on the ST and the MT output for this purpose. Among the features they used are source-sentence length, frequency of n-grams (i.e. contiguous sequences of words), language model probabilities,[11] as well as correspondence probabilities between the ST and the MT output.

Specia, Turchi, Cancedda, Dymetman and Cristianini (2009) and Specia, Raj and Turchi (2010) used subjective effort ratings (in 1-4 or 1-5 scales) provided by post-editors at the sentence level as a proxy for MT quality in the process of developing quality estimation systems. Specia (2011) also used editing time and HTER for this purpose and evaluated the performance of different systems based on a PE task, i.e. by asking translators to post-edit sentences for which reference translations were not available and checking to see if higher levels of productivity were achieved for sentences receiving high quality-estimation scores. Indeed, higher productivity was observed for sentences estimated by these systems as being of high quality, with the best result being obtained for systems based on PE time, as opposed to HTER or subjective ratings. Regarding the textual features linked to PE effort in the process of

[11] A statistical language model is a tool that estimates the probability of a sequence of *n* words occurring in the language, information which is normally obtained from large-sized text corpora.

developing these systems, while percentage of nouns was amongst the indicators presenting a high correlation with editing time for English STs, the features with the highest predictive power for French STs were mostly based on language model probabilities (see footnote 11) (English-Spanish and French-English were the two language pairs used in the study). In regard to these results, it is worth noting that Specia's study is based on a single post-editor per language pair. This ideally calls for a confirmation of these findings, which is partly undertaken in this thesis in the context of the French-English language pair.

2.3.2.3. *Human Evaluation*

Despite the many advantages offered by automatic metrics, human judgments are still considered the gold standard in MT (and PE) evaluation. One of the methods most frequently used for a human evaluation of MT is the one proposed by the Linguistic Data Consortium (2005, in Callison-Burch, Fordyce, Koehn, Monz and Schroeder 2007). Based on this method, the raw MT output is evaluated at a sentence level according to scales of 'fluency' (linguistic quality) and 'adequacy' (faithfulness to the ST). More recently, new versions of these scales have been proposed in the context of the Translation Automation User Society's (TAUS) Dynamic Quality Evaluation Framework.[12] This framework provides a set of guidelines and tools for MT evaluation that can also be used to evaluate the post-edited text (TAUS 2013a).

The fluency scale proposed in the TAUS framework is as follows:

4 – Flawless
3 – Good
2 – Dis-fluent
1 – Incomprehensible

The adequacy scale (reflecting the amount of information in the ST also present in the translation) is as follows:

4 – Everything
3 – Most
2 – Little
1 – None

Previous experiments have shown that making use of scales of this kind is not unproblematic. Callison-Burch et al. (2007) observed high degrees of correlation between fluency and adequacy. According to them, this might indicate that judges find it difficult to distinguish between these two concepts. Also, the levels of inter-rater

[12] See https://evaluate.taus.net/ (accessed 19 February 2015).

reliability (i.e. agreement between judges) they obtained in their study were deemed below expectations, implying that judges might considerably differ from one another in their interpretation of ranks in the scales. As a way of addressing some of these problems, Callison-Burch et al. tried a different approach. They randomly selected translations and asked judges to intuitively rank them based on their level of quality. The authors observed an improvement of inter-rater reliability for the ranking strategy in comparison with fluency and adequacy scales – differently, for example, from research in interpreting, where little difference has been observed between similar evaluation methods in terms of assessment consistency (see Wu 2010, 125).

Despite the advantages of the ranking method in the context of MT, it is worth noting that this strategy is associated with a more complex evaluation design. Since judges should ideally not be presented with too many sentences to rank at a time, this method ideally requires a large number of ranking events. Also, partial rankings (i.e. rankings comprising just a random selection of post-edited versions for the same ST sentence) should ideally be combined into an overall ordering, which is not unproblematic – though see the solution proposed by Lopez (2012). In the present project, carrying out a human assessment of *all* post-edited versions produced in the study as well as of the MT output was necessary, which indeed would require a large number of assessing events if the ranking method was to be used – i.e. because it is impractical to ask judges to rank 29 sentences (the number of post-edited versions produced in the study plus the raw MT output) at one time. In view of this, the relative simplicity of fluency and adequacy scales seemed to outweigh the complexity of the ranking method. The evaluation was based on TAUS's tool and guidelines, which provide a recent framework for fluency and adequacy assessment, consistent with current professional practice (see section 3.3.6 for details).

2.4. Findings from Previous Research

As previously mentioned, the methods of text analysis and translation evaluation described above are frequently used in PE research as a way of checking for potential correlations involving behavioural aspects of the activity and characteristics of the ST and/or MT output. Previous findings obtained based on comparisons of this kind are presented below (sections 2.4.1 - 2.4.4) together with a brief account of findings obtained in qualitative studies (section 2.4.5). A summary of these findings and a discussion on desiderata in previous research are provided at the end of the chapter (section 2.5).

2.4.1. *Source-Text Characteristics and Post-Editing Effort*

As a by-product of the main questions addressed in his study, Krings (2001) provided preliminary indications of how the ST can influence processing speed (number of ST words processed per time unit), verbalisation effort (number of words verbalised, under a think-aloud condition, per ST word), as well as the frequency of cognitive processes and attentional shifts in PE (calculated based on TAPs). Krings observed that longer STs were in fact associated with higher processing speed (i.e. less time per word) and less verbalisation effort. As an explanation for this arguably counter-intuitive finding, Krings suggested that the amount of time that goes into getting familiar with the text would be expected to have a higher relative impact on speed in the context of shorter STs. This is because once post-editors pass this familiarisation stage they would be faster in dealing with the remainder of the task, with this initial phase taking up a smaller proportion of total task time in the context of a longer text (ibid., 282). As for the cognitive processes associated with the ST, Krings attributed most differences in this respect to the texts' level of comprehensibility, estimated based on the number of inferential ST-related processes observed in the think-aloud data. In Krings's study, particularly slow processing speed was observed for a text showing a large number of such processes (ibid., 283).

While in terms of effort prediction Krings's findings reveal important information, he only considered the entire text as a unit of analysis whereas, as pointed out by Tatsumi (2010), analysing sentence-level ST characteristics may constitute a more precise approach in predicting PE effort. In addition, any ST features considered by Krings are based mainly on text and sentence length, characteristics that are likely to miss aspects of the texts lying at deeper linguistic levels. In response to this problem, potentially richer features such as lexical frequency and Coh-Metrix indices (see sections 2.3.1 and 3.4.1) are used in the present study.

Bernth and Gdaniec (2002) investigated ST characteristics and suggested a number of ways in which the machine translatability of the ST, i.e. how MT-friendly it is, can be enhanced by avoiding certain features such as ambiguity, coordination, ellipsis, etc. Features of this kind are usually referred to as negative translatability indicators (Underwood and Jongejan 2001) and have been particularly useful in the development of controlled languages, i.e. languages that are freed of these characteristics and therefore more likely to be translated correctly by MT systems. A number of previous studies looked at the relationship between controlled STs and PE effort, with good results normally being observed for the use of controlled language both in terms of amount of editing (Aikawa, Schwartz, King, Corston-Oliver and Lozano 2007) and processing speed (O'Brien 2006).

O'Brien (2004) analysed the relationship between negative translatability indicators and PE effort and found that, on average, the presence of these indicators in the ST decreases processing speed. She highlights, however, that different negative translatability indicators will have different levels of impact on effort. Abbreviations and proper nouns, for example, were amongst the features for which no effects were observed.

Tatsumi (2009) made one of the first attempts at statistically predicting PE time and showed that, in addition to scores reflecting amount of editing, features reflecting ST sentence complexity were able to improve the fit of regression models used in the analysis. Tatsumi and Roturier (2010) tested the correlation of PE time with other source features, such as scores produced with the Acrolinx IQ[13] and After the Deadline[14] tools, reflecting the quality of spelling, grammar and style. In Tatsumi and Roturier's study, only Acrolinx IQ presented a statistically significant effect out of all ST indices tested. Tatsumi (2010) also showed that a ST complexity score obtained through the MT system SYSTRAN correlated well with PE time. In regard to Tatsumi's (ibid.) findings, however, Green, Heer and Manning (2013) point out that the regression technique implemented does not account for variation between participants in the study and the wider population. In accounting for such variation, mixed-effects modelling, the technique adopted in the present thesis, is normally deemed more adequate (Balling 2008; Baayen, Davidson and Bates 2008) (see also section 3.4.5.2).

Green, Heer and Manning (2013) automatically parsed English STs with Stanford CoreNLP[15] to obtain features that could act as potential predictors of PE time. A significant effect was observed for the proportion of nouns in the ST, which was associated with more PE time per sentence. Green, Heer and Manning also exploited mouse hover patterns as an indicator of post-editors' attentional focus and found that both nouns and adjectives were frequently hovered over with the mouse.

Aziz, Koponen and Specia (2014) made use of (text) 'production units' to test the power of different ST features in predicting PE time and amount of editing. In their

[13] See http://www.acrolinx.com/acronews_en/items/acrolinx-iq-21-available-now.html (accessed 08 July 2012).

[14] See http://www.afterthedeadline.com/ (accessed 08 July 2012).

[15] This is a suite of natural language processing tools that tags and analyses text providing information such as part-of-speech categories and syntactic roles – see http://nlp.stanford.edu/software/corenlp.shtml (accessed 14 February 2015).

study, production units consisted of clusters of overlapping editing interventions (i.e. interventions corresponding to the same parts or segments of the ST) (ibid., 185). Their results indicated that production units involving verbs tend to be associated with more PE time, while production units involving nouns tend to be associated with more editing operations.

In regard to the results reviewed above, it should be pointed out that the majority of previous research investigating the impact of ST characteristics on PE behaviour has English as source language (e.g. Aziz, Koponen and Specia 2014; Green, Heer and Manning 2013). In this respect, it cannot be excluded that ST linguistic features may not generalise across different language pairs. This might be especially so in the case of features based on part-of-speech categories as these may play different roles depending on their linguistic context – e.g. nouns can directly modify other nouns in English, whereas in French this would normally require a preposition. In addition, most studies mentioned above do not take MT quality into account *together* with ST characteristics, an approach that does not allow distinguishing between features of the ST and of the MT output in terms of their potential for predicting PE effort (see section 2.5). In the present study, non-English STs are used and MT-output evaluation scores are taken into account together with ST features, a contrast not normally addressed in previous research.

2.4.2. *Machine-Translation Quality and Post-Editing Effort*

Few studies to date have examined potential correlations between MT quality measures and PE effort. Early work in this respect was carried out by Krings (2001). He collected human ratings of MT quality at a sentence level and observed that, contrary to expectations, sentences of medium quality were the ones that required the largest number of cognitive processes, postulating that these sentences required ‘a greater dispersion of attention across three different texts’ (ibid., 539), i.e. the ST, the raw MT output and the emerging TT, as mentioned in section 2.1.

O’Brien (2011) grouped ST-MT sentence pairs into three categories according to MT evaluation scores obtained with the GTM metric (Turian, Shen and Melamed 2003) (see section 2.3.2). She arbitrarily defined a ‘low’ band of the metric (expected to require high effort) as scores below 0.4 (inclusive) and a ‘high’ band (expected to require low effort) as scores above 0.81 (inclusive). Scores in between were defined as ‘medium’. O’Brien showed a linear relationship between these three categories and eye-tracking measures – namely, number of eye fixations landing on the text and average fixation duration (see section 2.2.3.1). She also checked to see how scores of the TER evaluation metric (see section 2.3.2) correlated with processing time (i.e. number of ST words processed per unit of time) and observed similar results as those obtained for the GTM metric. It is worth noting with regard to these findings that O'Brien's design required that sentence pairs be presented to post-editors in random order, breaking the original sequence of the ST. A setting of this kind cannot be excluded as interfering with cognitive processes during the task as, even for texts in the information technology domain (the subject matter adopted in her investigation), it would be expected that sentences in the same text bear semantic/syntactic links with each other, so post-editing these sentences in random order might require additional cognitive effort.

Gaspari, Toral, Naskar, Groves and Way (2014) examined potential correlations between PE time and three automatic evaluation metrics: BLEU, TER and Meteor. They observed weak to moderate correlations between PE time and scores of these three metrics (applied to entire documents as opposed to sentences). They also compared these results between PE and traditional translation (using the translations produced from scratch as the references – see section 2.3.2.1) and found more consistent correlations between these three metrics and time spent translating, as opposed to time spent post-editing.

Rather than making use of evaluation scores, some previous studies have focused on the nature and quantity of errors in the MT output as potential predictors of effort. Temnikova (2010) proposed a ranking of MT errors based on the amounts of cognitive effort that correcting these errors is expected to require. The main factors motivating this ranking are the linguistic level associated with the error (e.g. lexical or syntactical) and processing span, i.e. how far the area involving the error stretches in the text, where local problems are deemed easier to handle than problems stretching across longer spans. The ranking proposed by Temnikova (ibid., 3488) is presented below, from the least (1) to the most (10) cognitively demanding category:

1 – Correct word, incorrect form
2 – Incorrect style synonym
3 – Incorrect word
4 – Extra word
5 – Missing word
6 – Idiomatic expression

7 – Wrong punctuation

8 – Missing punctuation

9 – Word order at word level

10 – Word order at phrase level

As can be seen above, errors at more local linguistic levels, such as those concerning morphology and lexis, are regarded by Temnikova (2010) as less cognitively demanding to correct, ranking at categories 1-4. Errors of a syntactic nature, on the other hand, are deemed more demanding, ranking at categories 7-10. This ranking has been empirically tested by Koponen et al. (2012) in view of the amount of time post-editors take to post-edit sentences with these errors. Based on the results obtained, Koponen et al. proposed a number of adaptations to the ranking, including the addition of a 'zero' category reflecting mere typographical changes such as the use of lower/upper case, as well as a breakdown of category 3 ('incorrect word') into three sub-categories: (3a) different word but same part-of-speech; (3b) different part-of-speech; and (3c) untranslated source word in MT output. Koponen et al. also observed that issues involving punctuation (categories 7-8) were less cognitively demanding than estimated by Temnikova.

Lacruz, Denkowski and Lavie (2014) also investigated MT errors and their connection with PE effort. They looked at how error categories described in the American Translators Association assessment rubric (Koby and Champe 2013) were linked to HTER, subjective ratings on translation quality and average measures of pause time – namely, a measure of pause to word ratio, shown in previous work by Lacruz and Shreve (2014) to reflect cognitive effort. The highest correlations between MT errors and pause to word ratio observed by Lacruz, Denkowski and Lavie involved errors relating to terminology as well as content omission or addition. These results should be interpreted with caution, however, as individual sentences post-edited by the same participant were assumed as independent observations in the analysis – whereas there is the risk that differences between participants might, alone or in part, drive overall patterns (see e.g. Bland and Altman 1995) (see also section 3.4.5.2).

To the present author's knowledge, further information on potential connections between MT quality and PE effort can only be found in studies that do not explore MT quality itself, but rather the amount of editing implemented. This is the case of the study carried out by Tatsumi (2010), for example, who found that scores reflecting edit distance are significantly correlated with PE time. However, since Tatsumi instructed post-editors to carry out light PE, not asking them to render the post-edited text ‘stylistically sophisticated’ (ibid., 82), it is debatable if these scores can reflect the quality of the MT output. This is because this instruction is bound to restrict edit distance and consequently the level of quality of the post-edited output, which also applies, for example, to Tatsumi (2009) and Tatsumi and Roturier (2010). In the present study, rather than using measures of edit distance, MT quality is approximated either based on Meteor, which takes semantic information into account when comparing the MT output with a reference translation, or via a human assessment of both fluency and adequacy – estimates that are potentially closer to reflecting direct aspects of the quality of the MT output.

2.4.3. *Post-Editing Effort and Post-Editors’ Characteristics*

Very little research to date has explored connections between post-editors’ characteristics and PE behaviour. De Almeida (2013) looked for a potential link between time spent post-editing and the extent of participants’ previous professional experience, considering groups of both post-editors and translators. Perhaps surprisingly, she found no significant correlations in this respect, either in the case of previous experience in translation or PE. De Almeida also analysed post-editors’ performance, making use of four evaluation categories: ‘essential changes’, ‘preferential changes’, ‘essential changes not implemented’ and ‘introduced errors’ (ibid., 100). A number of sub-categories were also proposed, covering aspects pertaining, for example, to consistency and the use of pronouns and proper names. De Almeida concluded that the relationship between professional experience (in either translation or PE) and PE performance is an extremely complex one. In particular, the participants showing the best performance in her study did not have the highest levels of professional experience in the sample, leading her to suggest that, together with professional experience, other aspects of PE have a bearing on performance in the activity, such as for example participants’ attitude towards MT and their ability to adhere to the task brief.

Similarly to de Almeida (2013), Guerberof (2014) found no significant connection between amount of professional experience and PE time. In Guerberof’s study, however, when contrasting the extent of participants’ professional experience with both task time and post-edited quality, it was found that more experienced post-editors took less time to produce edited texts that were at the same level of quality as those produced by less experienced post-editors, potentially denoting that high levels of experience do not necessarily account for higher quality, but rather for higher efficiency.

In appraising the efficacy of PE relative to traditional translation, a number of variables pertaining to participants' individual characteristics were taken into account by Green, Heer and Manning (2013), including results of translation skills tests and the hourly rate participants in the study charged on the online website through which they were recruited. Green, Heer and Manning's experiments involved three language pairs: English into Arabic, German and French. Though small language-pair-specific effects were observed for participant variables, when taking all language pairs into account, no overall effects were found for any of the participant variables considered.

2.4.4. *Post-Editing Effort and Post-Edited Quality*

To the present author's knowledge, no study to date has directly explored potential connections between cognitive and other types of effort and the quality of the post-edited text, even though similar connections are tangentially touched upon in the context of investigations with different main objectives. In comparing PE to traditional translation, for example, Carl et al. (2011) observed that the amount of changes implemented in the MT output was not connected with subjective assessments on post-edited quality.

Green et al. (2014) were interested in examining the feasibility of interactive MT, an editing feature whereby machine translations adapt themselves on the fly to the text typed by the post-editor. In appraising this feature, they observed that PE time was negatively correlated with the final quality of the text. They regarded this result as an artefact of the evaluation score used to measure final quality, namely BLEU+1 (Lin and Och 2004), a sentence-level version of BLEU (Papineni et al. 2002). Since BLEU scores are known to favour short sentences (see section 2.3.2.1), Green et al. interpreted this result as a consequence of the fact that PE time was positively correlated with sentence length in their study.

Potential links between process behaviour and product quality have been more directly researched in traditional translation, where the usual hypothesis is that the more time and effort one invests in a translation, the better this translation will be (see e.g. Jääskeläinen 1996). A recent study, however, has come across a different result. Sun and Shreve (2014) asked 102 translation students to translate six short texts and found no significant correlation between translating time and translation quality. Similarly to the study conducted by Lacruz, Denkowski and

Lavie (2014), however, results reported by Sun and Shreve are based on correlation tests that assumed texts translated by the same participants as independent observations, a situation which, as previously mentioned, may fail to account for statistically harmful variation between participants (see Bland and Altman 1995).

2.4.5. *Qualitative Studies on Post-Editing*

Few qualitative investigations into post-editors' behaviour have been carried out in previous research. Krings (2001) proposed the most detailed description of PE operations available, with these operations being grouped according to cognitive processing categories presented in section 2.1. Krings catalogued and quantified the cognitive processes that take place in PE as a way of estimating cognitive effort. He was interested mainly in sequences of events that reflect *how* post-editors approached the task. This differs from the qualitative approach adopted in this thesis in two ways. First, in view of recent methodological criticisms, think-aloud data is not assumed in the present study a priori as an indicator of cognitive effort. Rather, potential correlations between TAPs and other measures of cognitive effort are empirically tested (see section 2.2.3.3, above). Second, the coding categories adopted by Krings were aimed at providing an operational description of PE processes rather than of the different types of issues and overall linguistic aspects of the task post-editors focus on, which is one of the research aims pursued here (see section 3.1.1).

Regarding other qualitative initiatives, Blain, Senellart, Schwenk, Plitt and Roturier (2011) presented a methodology for automatically classifying PE operations. The classification they propose is based on the logical links that bind together mere mechanical edits (deletions, insertions, substitutions and shifts). Blain et al. refer to logical units formed by these edits as PE actions. They relied on previous error-analysis literature to propose a typology of PE actions such as 'determiner choice', 'noun meaning choice', 'noun-phrase structure change', 'verb stylistic change', among others. By automatically tagging these actions, the system put forth by Blain et al. is able to produce detailed PE reports that can be used, for example, to improve MT systems by identifying repetitive actions that can be automatically propagated as a way of reducing PE effort. As a weakness of this strategy, however, the authors highlight that their findings and the automatic classification system proposed cannot be applied to a full-PE scenario (i.e. when final quality should be indistinguishable from human translation) or to situations that require post-editors to be creative. This is because these contexts would involve actions that are hard to predict, making it difficult for the system to identify editing patterns.

One of the research questions addressed by Tatsumi (2010) concerned the impact of different PE operations on PE speed. To identify these operations, she conducted a qualitative analysis of the PE process and proposed a typology of the different types of edits it involves. A summary of these edits (see Tatsumi 2010, 97ff.) is presented below:

Supplementation: when morphs[16] not present in the MT output are added to the final text to supplement or clarify meaning.

Omission: when information in the MT output is deleted.

User Interface Term Alteration: in the context of translating software documentation, this consists of altering terms that need to meet specific product specifications, such as menu items and messages.

Technical Term Alteration: when non-product-specific technical terms need to be altered in order to conform to company or TT conventions.

General Term Alteration: when lexical items (in general) are modified.

Bunsetsu Level Stylistic Change: *Bunsetsu* is a Japanese textual unit described by Tatsumi as consisting of one content word/morph accompanied by its function dependants. In the context of Tatsumi's analysis, this operation referred to changes at the *Bunsetsu* level that did not incur a change of meaning.

Dependency Edit: when dependency relations between *Bunsetsus* were modified.

Rewrite: when meaning modifications above the *Bunsetsu* level were carried out.

Structure Level Stylistic Change: when modifications that did not change meaning were carried out above the *Bunsetsu* level.

Similarly to Krings's (2001), Tatsumi's (2010) typology reflects *how* different PE operations are performed. In Tatsumi's case, however, only edits actually implemented are taken into account. Krings's use of think-aloud data arguably presents an advantage in this respect as it allows for edits or issues that are only mentally considered to be included in the analysis. Tatsumi's typology is also less generic than Krings's, focusing specifically on Japanese and on software documentation.

2.5. Summary of Previous Research and Desiderata in Related Work

Many factors would be expected to influence the expenditure of effort in PE, including characteristics of the ST, of the MT output and of participants themselves. Findings in this respect that are of key relevance to the present study were reviewed above. These

[16] Tatsumi uses the term 'morph' as the 'Japanese equivalent of 'word'' (Tatsumi 2010, 43).

include the fact that the incidence of nouns in English STs was found to be correlated with PE time (Green, Heer and Manning 2013; Specia 2011), that certain MT automatic evaluation metrics presented links with eye-tracking data (O'Brien 2011) and PE time (Gaspari et al. 2014), and that scores reflecting amount of editing were largely linked to PE time (Tatsumi 2010). In the context of this thesis, these findings have influenced, for example, the choice of STs to be used in the investigation, where the percentage of nouns and adjectives in different potential texts was calculated with a view to checking for any effects of these features in the context of French, the source language adopted for the study (see section 3.3.2).

More generally, the findings reviewed above are also useful in pointing to new avenues that deserve to be explored in empirical PE research. In the case of textual characteristics, through separate studies it is clear from previous research that both the ST and the MT output can influence the amount of effort expended by post-editors. However, it seems fair to assume that ST complexity and MT quality do not necessarily correlate in terms of the amount of PE effort they entail: the ST may be relatively easy to handle but the MT output of poor quality, or the MT output may be of high quality but the ST intrinsically complex. It can be hypothesised then that a comprehensive perspective on PE effort should take into account aspects relating to the ST *and* to the MT output as well as to post-editors' individual characteristics. To the present author's knowledge, this thesis constitutes the first attempt to jointly inspect the impact that these elements might have on behavioural aspects of the PE process.

As for the potential impact of post-editors' characteristics on the expenditure of effort, it could be argued that effects of this kind should be taken into account in *any* analysis based on human behavioural aspects of the activity and not only in studies whose main objectives involve an investigation of post-editors' traits. Here, a failure to properly handle individual differences between participants and/or differences between participant samples and the population may well have marred empirical results provided by a number of previous projects, including studies where the interdependence of observations stemming from the same post-editor is neglected. This problem is combatted in this thesis with the use of mixed-effects modelling, which is arguably the most appropriate method for analysing datasets where repeated observations (e.g. sentences or texts) are associated with the same subject. This is also a more rigorous method when attempting to extrapolate results based on a sample to the wider population (see section 3.4.5.2).

In addition to checking for correlations between PE effort and other elements involved in the activity, a smaller number of studies also set out to describe the PE process and the different operations it entails (e.g. Krings 2001; Blain et al. 2011; Tatsumi 2010). Here, most previous research relies on physical editing interventions, failing to cover less evident information such as ideas that are just mentally considered. The only exception in this respect is the study carried out by Krings (2001), who catalogued the different cognitive processes that take place in PE, providing the present thesis with a useful framework of the psycholinguistic operations PE is expected to involve (see section 2.1). As previously mentioned, however, Krings's is a different angle of analysis from the one adopted in this thesis. In the present study, general linguistic aspects of the task post-editors attend to are examined as opposed to the ways in which the task is carried out.

Finally, it is also worth noting that while a great deal of previous research focuses on comparing the economic feasibility of PE and traditional translation in terms of effort and product quality (e.g. Green, Heer and Manning 2013, Plitt and Masselot 2010), the connections between PE effort and the quality of the post-edited text has been touched upon to a far lesser extent (e.g. Carl et al. 2011; Green et al. 2014). As this is a topic that has the potential to reveal better ways of carrying out the activity, a more complete understanding of this connection is arguably central to the field. This thesis constitutes, to the knowledge of the author, the first study where different behavioural aspects of PE (e.g. cognitive effort, amount of changes implemented and total PE time) are taken into account as potential predictors of product quality whilst controlling for the quality of the raw MT output and for post-editors' different profiles.

Chapter 3. Methodology

This chapter describes the overall methodological strategy adopted in the study. Research aims and specific research questions are outlined in section 3.1. Overall considerations regarding the mixed-method approach adopted for the investigation are presented in section 3.2. The design of the study is described throughout section 3.3, including information on piloting stages (section 3.3.1), the materials used (sections 3.3.2 and 3.3.3) and participants (section 3.3.5). Data processing stages are outlined in section 3.4. A summary is provided in section 3.5.

3.1. Aims and Research Questions

3.1.1. *Research Aims*

The overall purpose of the present study is to investigate the expenditure of cognitive effort in PE (see section 1.1). As mentioned in section 2.2.1.1, Krings (2001) defines overall PE effort based on three underlying variables: temporal, technical and cognitive effort. Of these, Krings regards *cognitive* effort as the main factor influencing the overall expenditure of effort in PE. As discussed in section 1.1, temporal measures or number of editing operations alone may fail to reflect the cognitive complexity of the post-editing activity (de Almeida 2013; Koponen 2012). Nevertheless, most previous research takes just these two parameters into account, contrasting them either with ST complexity or MT-output quality (see sections 2.4.1 and 2.4.2), but not with both, which could be seen as an incomplete approach to PE effort prediction (see section 2.5). In response to this, a first aim of the present study consists of identifying, as far as possible, what different elements involved in PE seem to have a connection with *cognitive* effort (as defined in section 2.2.1.3), with characteristics of the ST, MT output, post-editors, or the task itself being tested as potential predictors.

A second aim of the study consists of exploring the relationship between the effort invested in the activity (cognitive as well as other types) and the resulting levels of quality of the final post-edited text. Here, most previous research compares levels of quality achieved through PE and through traditional translation, with very little information being available on the relationship between PE behaviour and product quality; information that has direct applications for professional practice and which can further the current understanding of the activity.

In pursuing these two aims, different types of data are exploited in the investigation; in particular, subjective ratings, eye movements and TAPs (think-aloud protocols). Of these, TAPs are the data type with the least established tradition as a measure of cognitive effort, having been a target of criticism in translation process research due to potential validity and interference issues (see section 2.2.3.3). In view of this, from a methodological perspective, a third aim of the study consists of checking to see, in the context of PE, to what extent any links can be observed between the various data types used in the investigation to estimate cognitive effort and observe the PE process.

In summary, the three aims of the study are as follows:

Aim 1: Investigating the post-editing process with a view to identifying links between *cognitive effort* and elements pertaining to the *ST*, the *MT output*, *post-editors* and the *task itself*.

Aim 2: Investigating the relationship between (1) the levels of *cognitive and other types of effort* invested in the task and (2) the *quality of the results obtained*.

Aim 3: Investigating the relationship between *different types of data* that may be used as *measures of cognitive effort* in PE.

Specific research questions deriving from these aims are outlined below.

3.1.2. *Research Questions*

Cognitive effort is estimated here based on eye-tracking metrics as well as on subjective ratings (see section 3.4.3). As mentioned in section 2.2.1.2, cognitive effort is a complex construct which can be influenced by demands that are intrinsic to the task as well as by characteristics of the individuals that carry it out. With this in mind, in the context of Aim 1 cognitive effort is contrasted with potential predictors including automatic MT evaluation metrics, linguistic and psycholinguistic features deemed to reflect the complexity of both the ST and the MT output, and subjects' individual characteristics, namely working memory capacity, level of proficiency in the source language and amount of professional experience (in PE, translation or translation revision). From a broader perspective, different linguistic aspects of the task post-editors attend to are also considered, such as grammar, lexis, and adequacy to readership and translation context, aspects which are investigated both in their own right as well as in view of their potential connection with cognitive effort. Research questions (RQ) deriving from the first aim of the investigation are as follows:

RQ1: What *textual characteristics* pertaining to the ST or the MT output seem to bear a relationship with the amount of cognitive effort expended by post-editors?

RQ2: What *individual characteristics* seem to bear a relationship with the amount of cognitive effort expended by post-editors?

RQ3: What is the nature and frequency of different *linguistic aspects of the PE task* post-editors attend to?

RQ4: What *linguistic aspects of the PE task* seem to bear a relationship with the amount of *cognitive effort* expended by post-editors?

In the context of Aim 2, post-edited quality is human-assessed in terms of adequacy and fluency (see section 3.3.6), with both these variables being contrasted with different types of effort measures reflecting PE behaviour, such as amount of editing and eye fixations. Aim 2 is pursued here in attempting to answer the following specific question:

RQ5: What is the nature of the relationship between (1) *different types of effort invested in PE* and (2) *post-edited fluency and adequacy*?

In pursuing Aim 3, the present study examines how TAPs compare to eye movements and subjective ratings, data sources that have been more traditionally assumed as measures of cognitive effort in previous research. Aim 3 corresponds to the following specific question:

RQ6: Do *TAPs* correlate with *eye movements* and *subjective ratings* as *measures of cognitive effort* in the context of PE?

Vieira (2014) shows a preliminary analysis of RQ1 and RQ2, above, based on a subset of the participant sample used here. This analysis is expanded in this thesis.

3.2. A Mixed-Method Approach

Some of the questions outlined above are arguably better addressed based on quantitative evidence, such as RQ1 and RQ2 (i.e. the connections between, on the one hand, cognitive effort and, on the other hand, textual features and subjects' characteristics, respectively). Other questions were deemed to more suitably fit a partially qualitative approach, such as RQ3 and RQ4, where a qualitative analysis is able to reveal different linguistic aspects of the task post-editors deal with. In view of the mixed nature of the questions addressed and due to the arguable advantages of exploiting a variety of perspectives, a mixed-method approach (Creswell 2009) was adopted in the study. For this purpose, data is collected in the context of two tasks: one geared towards a quantitative analysis, based on eye tracking and subjective ratings (henceforth referred to as just 'the eye-tracking task'), and one geared towards a partially qualitative analysis, carried out under a think-aloud condition (henceforth 'the think-aloud task').

Two tasks were deemed necessary because collecting eye movements, subjective ratings and TAPs in the same task would invalidate any comparisons carried out in the context of RQ6 (i.e. links between TAPs and other types of data). This is because the slow-down effect induced by the think-aloud condition (see section 2.2.3.3) would inevitably influence eye-tracking and self-assessment data, i.e. by resulting in the recording of a larger amount of eye movements and by potentially also making participants perceive the task differently, altering their effort ratings.

In addition to allowing RQ6 to be addressed, the mixture of methodologies adopted here also allows pursuing the generally sought research aim of achieving both breadth and depth of analysis, which was attempted in this thesis by conducting tasks with a view to obtaining both quantitative and qualitative findings. In particular, results based on the eye-tracking task are in part exploited here as a way of informing the analysis carried out based on TAPs, by providing information on the different levels of cognitive effort that ST-MT sentence pairs sampled for the study were expected to pose (see section 4.2.3). A setting of this kind is in line with the 'sequential explanatory strategy' of mixed research design described by Creswell (2009, 211), where qualitative data obtained in a second research phase builds on quantitative results obtained in a first phase. Taking this into account, connections between results obtained in the eye-tracking and think-aloud tasks were established whenever possible, with findings arrived at in one task being exploited to expand and exemplify results obtained in the other.

As any research method will suffer from inevitable weaknesses, the combination of data from the two tasks conducted in the study allows the strengths of one method to compensate for the deficiencies of another. This notion is related to the principle of 'triangulation', a term used in the past in allusion to a search for convergence between qualitative and quantitative analyses, but which more recently has moved on to imply the actual integration of these two approaches (Creswell 2009, 14).

A scheme representing the overall design of the investigation and the specific research questions tackled by each task is presented below in Figure 3.

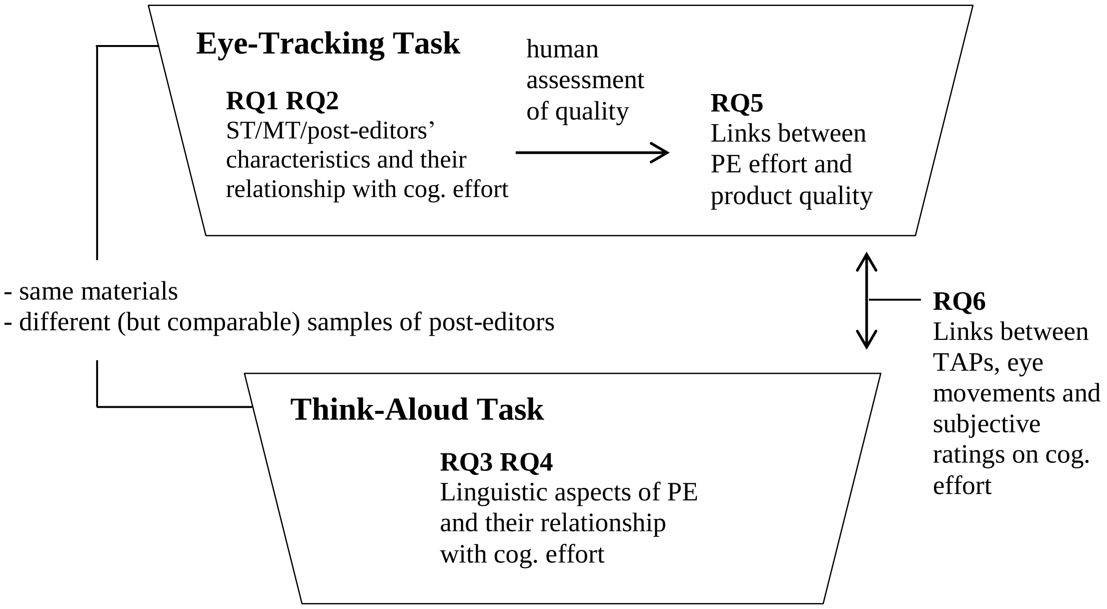

Eye-Tracking Task
RQ1 RQ2
ST/MT/post-editors' characteristics and their relationship with cog. effort
human assessment of quality
RQ5
Links between PE effort and product quality
- same materials
- different (but comparable) samples of post-editors
RQ6
Links between TAPs, eye movements and subjective ratings on cog. effort
Think-Aloud Task
RQ3 RQ4
Linguistic aspects of PE and their relationship with cog. effort

Figure 3 - Scheme of tasks and research questions of the study.

RQ1, RQ2 and RQ5 (i.e. effort vs. textual features, participants' characteristics and post-edited quality, respectively) are addressed primarily based on data obtained from the eye-tracking task, with RQ5 resting on a human assessment of both post-edited versions and the raw MT output (see section 3.3.6.1). Data collected in the think-aloud task is then analysed with a view to answering RQ3 and RQ4 (i.e. linguistic aspects of the task attended to and how these aspects are connected with effort). Edited versions produced in the think-aloud task are also assessed, but as the think-aloud condition is expected to influence product quality (Jakobsen 2003), this assessment acts as just a complement to the analysis carried out in the context of RQ5 and not as the main source of data used to answer the question. The same texts are edited in both tasks, by different, but comparable, samples of participants (see section 3.3.5). This allows a link to be established between eye-tracking and think-aloud results, a connection exploited when addressing RQ6 (comparison between potential indicators of cognitive effort) as well as in the process of answering other research questions. Specific information on the design of and procedure for conducting the tasks used in the study are provided below.

3.3. Research Design

3.3.1. *Piloting Stages*

The research questions addressed in this thesis pose a number of methodological challenges. Selecting STs that varied in complexity and translations that varied in quality was a preliminary requirement, since both these variables are compared here against cognitive effort. The text samples used for analysis should also ideally be relatively short so as not to take much of participants' time, which represented an additional difficulty in achieving the aforementioned variation.

A number of piloting stages were required for these challenges to be satisfactorily addressed. A pre-pilot study was initially carried out solely for the purpose of testing the suitability of different types of equipment and research-oriented text-editing platforms. For further information on this pre-piloting stage see Vieira (2013), which shows an evaluation of different tools that can be used for PE process research. Subsequent piloting stages are described below.

3.3.1.1. *Pilot Study I*

Three participants were recruited for a first pilot study aimed at testing different materials as well as different strategies for the selection of STs and machine translations. News was the text genre chosen for the investigation. A number of reasons motivated this decision. First, news texts can be relatively generic and their format is familiar to most people, not requiring extra genre-specific knowledge from participants, which could limit the size of the subject sample. Second, much of previous research on the development and evaluation of MT systems is based on the use of news texts as test-sets, resulting in a number of such texts and their corresponding machine translations being freely available for research purposes, which constituted rich material that could be exploited in the present investigation. Third, most research on PE to date has been carried out based on genres that are relatively MT-friendly, such as user manuals and software documentation (see Calude 2002). While genres of this kind are certainly in line with real-world professional practice they are also less linguistically varied. Because of this and in view of their non-technical nature, news texts would be expected to pose more challenging problems for MT, which, from a research perspective, seemed like an interesting feature to exploit.

Different news articles were tested in the first pilot study, with a number of adjustments in the design and choice of materials deriving from this initial stage. For example, one of the news articles initially considered was about the US financial market. While at first sight this text did not seem particularly difficult, being scored at a moderate level of readability (see section 3.3.2), participants in this first pilot study felt that a considerable level of subject-specific knowledge was required to post-edit this text. A situation of this kind would be hard to control because participants' knowledge of the subject matter would have too strong an influence on the nature of the results obtained. In view of this, it seemed desirable to select texts that were for the most part self-contained, not requiring profound knowledge of the topic from the readers.

Other modifications deriving from this first piloting phase included slight adjustments in the editing interface, clearer task instructions and longer warm-up tasks (see section 3.3.4).

3.3.1.2. *Pilot Study II*

A second pilot study was carried out with an additional four participants, now with the aim of testing the research design. In dealing with methodological challenges similar to the ones faced here, previous research has often opted for presenting texts for editing as separate sentences in random or non-sequential order (e.g. O'Brien 2011; Specia 2011). This was avoided in this thesis by mixing in the same text machine-translated sentences of different quality levels, produced by different MT systems, presented for editing in source-document order. The systems used were either generic commercial engines or in-domain systems tuned with news texts, the genre adopted for the study.

Results obtained from the second pilot study showed a significant correlation between the automatic translation evaluation metric Meteor (Denkowski and Lavie 2011) (see section 3.3.3) and the number of eye fixations landing on ST-MT sentence pairs[17] ($\rho(34) = -0.35$, $p < 0.05$). It was noticed, however, that by excluding data from the first three sentences each participant looked at, the strength and significance of the

[17] The number of fixations was normalised by the character count of each source sentence to cancel effects of sentence length. Spearman's *rho* (ρ) was the correlation test used as the data was not normally distributed. Participants edited two texts, in alternate order, and results are based on per-sentence averages.

correlation increased (ρ(34) = - 0.40, p = 0.01). Even though participants had a chance to carry out a warm-up task, this difference seemed to reflect an acclimatisation effect (see section 2.4.1), also observed in previous research (e.g. Doherty, O'Brien and Carl 2010; Krings 2001). In view of this, in the main eye-tracking task it was decided to disregard data relating to the beginning of the texts: three sentences in the case of one text and four in the case of another. These sentences consisted of the title plus an opening paragraph before the body of the texts[18] for which human reference translations were not available in the dataset with STs and MT outputs used in the study (see section 3.3.2, below).

Based on mixed-effects regression modelling (see section 3.4.5.2), the second pilot study also indicated a potential relationship between cognitive effort and the percentage of prepositions in the ST, an effect that was explored further in the main analysis.

Other adjustments deriving from the second pilot study include the use of slightly longer excerpts of the news articles selected and more time being dedicated to think-aloud warm-ups. It was also deemed desirable to widen the range of raw MT quality in the sample, which was achieved by selecting translations from a larger pool of systems. Specific details of the design of the main tasks are provided below between sections 3.3.2 and 3.3.4.

3.3.2. *Selection of Source Texts*

Excerpts of two news articles originally published in French were used in the study. The articles were retrieved from the *newstest2013* dataset, which results from 2013's Workshop on Statistical Machine Translation.[19] The selected texts are about prostate cancer screening (text A) and the voting system in the United States (text B) – see Appendix A.

Care was taken in choosing texts that did not have overly specific subject matters and which had different levels of translating difficulty, which would allow for more variance in ST features tested in the study at a sentence level. A number of textual features reported in previous research were exploited in the process of selecting the texts, including readability, word frequency and non-literalness – features that have been described as 'strong indicators of source-text difficulty in translation' (Hvelplund

[18] In journalism, this type of paragraph is known as 'standfirst'.

[19] This is an annual workshop where different MT systems are ranked according to human-assessed quality, with STs, MT outputs and human reference translations being available afterwards for research purposes.

2011, 88). Details of the selected texts are shown in Table 2. To give an indication of how these features varied within each text, they are presented both for the entire texts and for shorter passages. To allow for readability measurement (which requires text passages to be of a certain length – see section 2.3.1), each passage had at least 100 words and they also respected paragraph and sentence boundaries.

	LIX Readab.	KM Readab.	%N+Adj	Freq. (% on 1K list)	Non-lit. expression count	Word Count	Avg. sentence length (in chars., w/ white spaces)
PassageA1	45.6	58	32	79.3	1	109	107±71 SD
PassageA2	47.8	*68*	27	84.1	*2*	150	132±95 SD
PassageA3	46.2	61	21	82.3	0	135	121±40 SD
Text A	*47.8*	*63*	*26*	*82.2*	3	*394*	*122±67 SD*
PassageB4	56.3	48	44	75.2	*4*	172	128±33 SD
PassageB5	54.2	50	40	75.8	0	134	124±54 SD
PassageB6	55	54	34	81.5	1	144	147±50 SD
Text B	*55*	*51*	*39*	*77.4*	*5*	*450*	*132±44 SD*

In the LIX scale, higher values indicate less readability, while higher KM values indicate higher readability. Figures exclude titles and introductory paragraphs before the body of the texts.

Table 2 – Details of the estimated translating difficulty of the source texts selected for the study.

The readability metrics exploited are two formulae that can be used for French, namely LIX formula[20] (Björnsson 1968) and Kandel and Moles (KM)[21] (Kandel and Moles 1958) ('LIX Readab.' and 'KM Readab.' in Table 2, respectively).

Nouns and adjectives are reported by Green, Heer and Manning (2013) as showing an association with PE time and mouse hover patterns for English, so their combined percentage in the passages ('%N+Adj') was also taken into account, thus checking for any effects of these part-of-speech categories for French STs. Part-of-speech percentages were obtained by tagging the texts with Stanford Part-of-Speech Tagger (Toutanova, Klein, Manning and Singer 2003).

Lexical frequency ('Freq. (% on 1K list)') was calculated based on the percentage of words in the passages that could be found on a list of the one thousand ('1K') most frequent words in French (Jones 2000).

Non-literalness ('Non-lit expression count') was measured in terms of metaphoric expressions as well as French words that deviated from their most common

[20] LIX is a Swedish abbreviation for läsbarhetsindex (readability index). LIX was used here based on the online tool available at http://www.mancko.com/tests-de-lisibilite/fr/ (accessed January 2014).
[21] Based on the online tool available at http://www.standards-schmandards.com/exhibits/rix/ (accessed January 2014)

lexical sense as per entries of the Larousse French Dictionary online.[22] Words and expressions in the texts regarded here as non-literal are shown below in Table 3.

Text A	Text B
il y avait beaucoup de contamination entre les groups témoins 'there was a lot of contamination between witness groups' (in this context, *groups témoins* = test groups)	Les dirigeants *républicains* 'The Republican leaders' (members of the Republican party in the US, as opposed to favourable or relative to the republic)
la clé est de prendre 'the key is to get'	*la nécessité de lutter contre la fraude* 'the need to fight against fraud'
Il y a des cancers agressifs et d'autres qui sont indolents 'some cancers are aggressive and others are indolent' (in this context, *indolent* = benign)	*Le Centre Brennan considère* […] *comme un mythe* 'The Brennan Center considers […] a myth' (metaphoric comparison with 'myth')
	ces mesures mineront le système démocratique 'these measure will mine the democratic system'
	les personnes à faible revenu 'people with weak income'

Table 3 - Examples of non-literalness observed in STs selected for the study.

Out of all news articles in the *newstest2013* dataset that fit the criteria of not having culturally marked items or overly specific subject matters, the selected texts were the ones with the most discrepant percentages of nouns and adjectives combined (26% in the case of text A and 39% in the case of text B). Measures of readability, word frequency and non-literalness roughly follow this pattern, with text A receiving scores indicating higher readability, a larger proportion of frequent words and a smaller number of non-literal expressions (LIX: 47.8; KM: 63; 82.2% of words on 1K word list; 3 non-literal expressions), and text B receiving scores indicating lower readability, a smaller proportion of frequent words and a larger number of non-literal expressions (LIX: 55; KM: 51; 77.4% of words on 1K word list; 5 non-literal expressions), as can be observed in Table 2.

It should be noted at this point that the impact of ST features on cognitive effort is examined at a sentence level in the present study; the preliminary textual analysis reported above acted mainly as a way of obtaining rough estimates on the translating difficulty of the texts in an attempt to cover a wider range of scenarios, i.e. by exposing participants to STs which differed in translating difficulty (at the text level) and also as a way of ensuring higher variance amongst the source sentences, overall.

[22] See http://www.larousse.com/en/dictionaries/french/ (accessed January 2014).

By collating the STs in the *newstest2013* dataset with the original news articles published online it was noted that short introductory paragraphs after the title and, in one of the texts, two segments in the body of the article, could not be found in the dataset versions. To make sure participants did not miss potentially relevant information, sentences in these additional passages were included in the materials, and the texts were presented for editing starting from the title. Translations for these extra sentences were randomly selected from online and commercial MT systems (i.e. because they were not in the *newstest2013* dataset), but data corresponding to these sentences was not analysed in the eye-tracking task because (i) reference translations were not available for these sentences, which prevented the use of automatic evaluation metrics (see section 3.3.3), and (ii) to avoid acclimatisation effects, as mentioned previously in section 3.3.1.2. Overall, 1037 source words were presented for editing, with eye-tracking data being recorded and analysed for 844 source words and 41 sentences.

3.3.3. *Selection of MT Output*

The *newstest2013* dataset was used as the main source for the selection of the MT output. To increase variability in quality in the sample, in addition to outputs from systems already included this dataset, translations produced with a further three online and two commercial systems were selected, forming a corpus of 24 candidate translations (19 already in the dataset plus five newly harvested ones) for each source sentence.

Version 1.4 of the Meteor automatic evaluation metric (Denkowski and Lavie 2011), run at default settings, was used as the basis for the selection. Meteor measures the similarity between the translation being assessed (usually referred to as the ‘hypothesis’ translation – normally MT) and a translation regarded as reference (normally a human translation). Meteor differs from more traditional metrics, such as BLEU (Papineni et al. 2002), in that it takes semantic information such as synonymy and paraphrasing into account when making this comparison (see section 2.3.2). The Meteor scale varies between 0.0 and 1.0, where 1.0 represents the perfect match of a hypothesis (Hyp) with a reference (Ref). Examples of Meteor scoring are provided below.

Ref: In addition, five million new voters in 2012 do not have such identification.
Hyp 1: *In addition*, five million new voters in 2012 do not have such identification. (1.0)
Hyp 2: *What is more*, five million new voters in 2012 do not have such identification. (0.95)
Hyp 3: *In contrast*, five million new voters in 2012 do not have such identification. (0.53)

As can be seen above, Hyp 1 is a perfect match with the reference, thus receiving the Meteor score of 1. Hyp 2, though bearing a different term to the reference – 'what is more' instead of 'in addition' – receives a high score (0.95), which could be explained by the semantic proximity between the two sentences. Hyp 3, in turn, modifies the meaning of the reference with the use of 'in contrast'. Hyp 3 receives the comparatively lower score of 0.53. In view of this capability of taking synonyms into account, it could be argued that semantics-based evaluation metrics are more sensitive to the accuracy of the MT output, which might not be the case for traditional metrics that calculate similarity (between MT and reference translation) based only on identical words.

The human reference translations used here are those included in the *newstest2013* dataset, produced by professional translators (see Bojar et al. 2013). For the selection of machine-translated sentences, Meteor scores were considered within one-decile 'ranges', with scores from 0.0 to 0.09 (inclusive) representing the first (bottom) range, and scores from 0.9 to 1.0 (inclusive) representing the tenth (top) range. Of ten possible ranges, 0.0-0.09 and 0.7-0.79 were absent from the MT output materials. Translations in these ranges, however, account for approximately 1% of all 57K sentences in the FR-EN *newstest2013* dataset, so the absence of these scores from the study was not deemed a sampling problem, but rather a scoring tendency of Meteor.

There were 24 translations available for each source sentence. At a first step, a random selection was carried out so that each source sentence was associated with one translation per Meteor range. To arrive at one MT version for each source sentence, translations were then selected step by step so as to obtain the best possible balance of Meteor values in the sample.[23] As there were 41 source sentences and 8 Meteor ranges available, a balanced distribution of Meteor values in the study would be approximately five sentences per Meteor range. This, however, was not possible as there were fewer than five sentences in certain ranges, so translations within under-represented ranges of

[23] Systems in the *newstest2013* dataset which entered the study sample are 'Online-A', 'Online-B', 'Online-G', 'CU-Zeman', 'FDA', 'CMU-syntax', 'rmbt-3', 'rmbt-4', 'KIT', 'DCU-FR-EN-Primary', 'MES-Simplified', 'Edinburgh' and 'Edinburgh-unconstrained'. Additional systems not in the dataset that entered the sample are SDL Freetranslation.com (http://www.freetranslation.com), TransPerfect (http://web.transperfect.com/free-translations/) and Microsoft Translator, the latter via MS Word (all harvested in October 2013).

Meteor were given priority and included in the sample directly, with the remaining sentences being randomly selected.

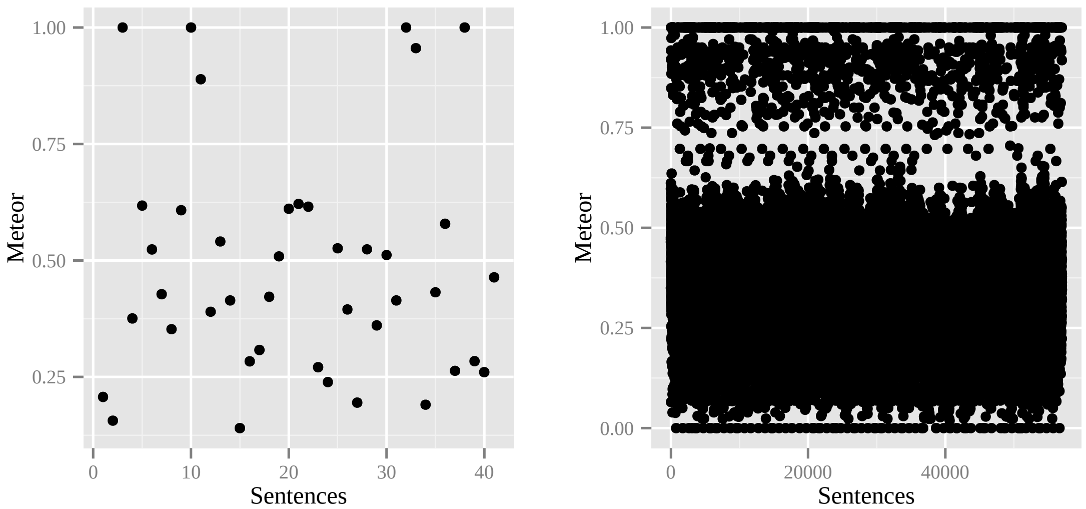


Figure 4 - Distributions of Meteor scores in the study sample (left) and in the entire FR-EN *newstest2013* dataset.

The distribution of the resulting selection can be observed in Figure 4. As can be seen, the distribution of the study sample is comparable to the distribution of the entire FR-EN *newstest2013* dataset. By adopting such a comprehensive approach, translations with scores as low as 0.14 were present in the materials. Upon visual inspection it was noted that translations below 0.20 were of extremely poor quality, including non-translated and malformed words.[24] A decision had to be made in this respect as to whether to keep these sentences or clean the sample. Meteor is one of the features being tested in the eye-tracking task, so having the entire spectrum of the metric seemed desirable. In view of this, these sentences were kept in the first instance, but further tests were carried out during data analysis by excluding from the sample machine translations with low human assessed scores to check for potential outlier effects (see next chapter).

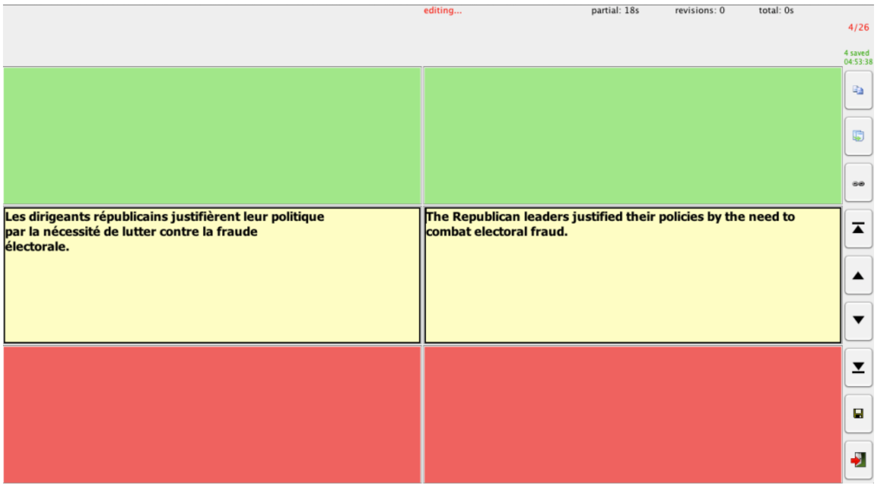
3.3.4. *Conducting the Tasks*

From a quantitative perspective, a sentence-by-sentence analysis of the data is desirable as this allows for more statistical power and more fine-grained results without needing to ask participants to post-edit a large number of texts. Collecting eye-tracking data for individual sentences when the whole text is displayed on screen is, however, a complex

[24] One of the systems used seemed to have a pre-processing problem, with French accented characters not being properly handled. This, however, was not a consistent pattern.

undertaking. This ideally requires the use of gaze-to-word mapping algorithms capable of automatically estimating on which letters/words participants' gaze lands. As these algorithms are still relatively fuzzy – their accuracy being reported at 65-88% in previous research (Jensen 2008; Dragsted and Hansen 2008, in Dragsted and Carl 2013) – it seemed desirable to design the eye-tracking task with participants being exposed to one sentence at time, with backtracking not being allowed. This renders eye-tracking data more reliable as participants' eye fixations can only pertain to each ST-MT sentence pair being displayed. However, a setting of this kind deviates from professional practice, where post-editors would normally have access to the entire text. In response to this, as the think-aloud task does not rely on eye tracking, participants were exposed to the whole text at once in that task, being allowed to move backwards and forwards in the text, under more realistic conditions. By comparing data in these two tasks, which are based on the same texts and on different, but comparable, samples of participants (see section 3.3.5), it was possible to estimate if the restrictive operating conditions of the eye-tracking task had a significant impact on results, which is briefly addressed in section 4.3.2.

PET (Aziz, Castilho and Specia 2012) was the editing platform adopted for the eye-tracking task. After confirming each sentence, participants were also prompted to provide subjective ratings on cognitive effort by using a scale (see section 3.4.3) configured within PET's interface. Figure 5 shows PET's editing interface as set up for the task.

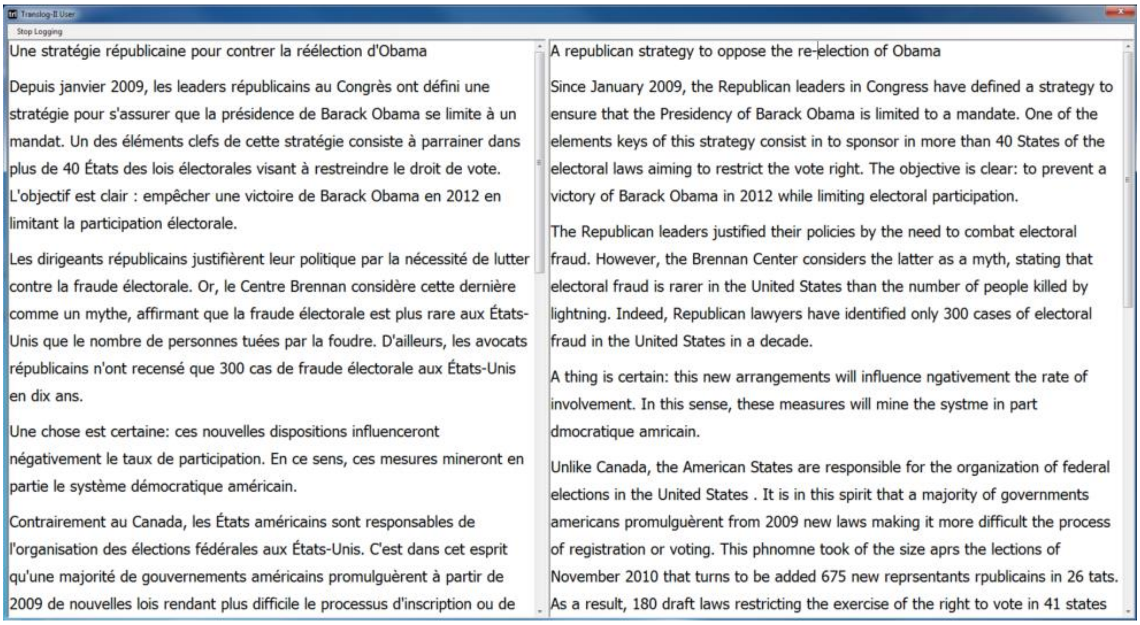


Les dirigeants républicains justifièrent leur politique
par la nécessité de lutter contre la fraude
électorale.

The Republican leaders justified their policies by the need to
combat electoral fraud.

Figure 5 - PET's interface as set up for the study.

The think-aloud task was carried out in Translog-II (Carl 2012), a tool deemed here more suitable for a setting where participants would have access to the whole text and which also produces linear key-logging reports which could serve as a support for the analysis of TAPs. The interface used in the think-aloud task can be observed in Figure 6, with the French ST on the left and the MT output on the right.



Une stratégie républicaine pour contrer la réélection d'Obama

Depuis janvier 2009, les leaders républicains au Congrès ont défini une stratégie pour s'assurer que la présidence de Barack Obama se limite à un mandat. Un des éléments clefs de cette stratégie consiste à parrainer dans plus de 40 États des lois électorales visant à restreindre le droit de vote. L'objectif est clair : empêcher une victoire de Barack Obama en 2012 en limitant la participation électorale.

Les dirigeants républicains justifièrent leur politique par la nécessité de lutter contre la fraude électorale. Or, le Centre Brennan considère cette dernière comme un mythe, affirmant que la fraude électorale est plus rare aux États-Unis que le nombre de personnes tuées par la foudre. D'ailleurs, les avocats républicains n'ont recensé que 300 cas de fraude électorale aux États-Unis en dix ans.

Une chose est certaine: ces nouvelles dispositions influenceront négativement le taux de participation. En ce sens, ces mesures mineront en partie le système démocratique américain.

Contrairement au Canada, les États américains sont responsables de l'organisation des élections fédérales aux États-Unis. C'est dans cet esprit qu'une majorité de gouvernements américains promulguèrent à partir de 2009 de nouvelles lois rendant plus difficile le processus d'inscription ou de

A republican strategy to oppose the re-election of Obama

Since January 2009, the Republican leaders in Congress have defined a strategy to ensure that the Presidency of Barack Obama is limited to a mandate. One of the elements keys of this strategy consist in to sponsor in more than 40 States of the electoral laws aiming to restrict the vote right. The objective is clear: to prevent a victory of Barack Obama in 2012 while limiting electoral participation.

The Republican leaders justified their policies by the need to combat electoral fraud. However, the Brennan Center considers the latter as a myth, stating that electoral fraud is rarer in the United States than the number of people killed by lightning. Indeed, Republican lawyers have identified only 300 cases of electoral fraud in the United States in a decade.

A thing is certain: this new arrangements will influence ngativement the rate of involvement. In this sense, these measures will mine the systme in part dmocratique amricain.

Unlike Canada, the American States are responsible for the organization of federal elections in the United States . It is in this spirit that a majority of governments americans promulguèrent from 2009 new laws making it more difficult the process of registration or voting. This phnomne took of the size aprs the lections of November 2010 that turns to be added 675 new reprsentants rpublicains in 26 tats. As a result, 180 draft laws restricting the exercise of the right to vote in 41 states

Figure 6 - Translog-II’s interface as set-up for the study.

Warm-up tasks were conducted prior to both the eye-tracking and think-aloud tasks to acquaint participants with the set-up and the think-aloud condition. In the eye-tracking task, the warm-up phase was based on the editing of practical task instructions, such as how to make use of buttons on PET’s interface and the fact that keyboard shortcuts such as Ctrl+C and Ctrl+V could be used. In the think-aloud task, the warm-up phase involved editing task instructions as well as a short excerpt of a news article also taken from the *newstest2013* dataset, all under a think-aloud condition. In both warm-up tasks, translations for the French STs were sampled from the same online/commercial MT systems used in the main tasks.

The order of presentation of texts (A or B) was alternated between participants in both eye-tracking and think-aloud tasks, with a break in between. Before the tasks participants were asked to consider an editing brief with instructions that included higher-quality PE guidelines put forth by the Translation Automation User Society (TAUS) (TAUS/CNGL 2010) – see Appendix B. They were told to implement as few changes as necessary to render the post-edited news articles as close as possible to being adequate for publication in an English-speaking context. No time limit was imposed in either task, but participants were told to attempt to finish the tasks in as little time as possible.

The constraints posed by the use of eye tracking also required that participants not be allowed to consult reference materials during the eye-tracking task. This is because, as pointed out by Hvelplund (2011, 86), the use of reference materials could lead to losses of eye-tracking data or a considerably more complex analysis, as fixations corresponding to reference-search events would ideally need to be identified and taken into account. Even though eye tracking is not the main method of data collection used in the think-aloud task, this restriction was maintained in both tasks for consistency.

Before editing each article, participants were asked to read a short text with factual background information on the subject matter of the texts to be edited (these are available in Appendix C). Even though the selected texts did not require overly specific extra-linguistic knowledge from participants, this seemed desirable as a way of levelling any potential discrepancies between participants in terms of previous knowledge of the subject matter and also as a way of tentatively alleviating the restriction on the use of reference materials. Participants were also told to trust the MT output when they did not know, and could not infer, the meaning of words in the ST.

It seemed desirable to gather as much background information about the participants as possible, as this would allow RQ2 (i.e. links between cognitive effort and participants' characteristics) to be addressed as well as allow these variables to be controlled in the analysis. To this end, after post-editing the texts, participants' working memory capacity, as defined in section 2.2.2.1, and level of French (the source language) were measured in two additional tasks. Participants' knowledge of French would naturally be expected to influence their editing behaviour as well as the amount of effort they expend, which may be especially so in view of the restriction on the use of reference sources. As for their working memory capacity, this seemed like a desirable construct to estimate as a number of previous findings show positive effects of working memory capacity on the performance of complex information-processing tasks (see section 2.2.2.1).

The working memory capacity of participants taking part in the think-aloud task was not measured. This was a way of taking less of these participants' time as this task proved to be slightly longer – due to the slow-down effect of TAPs (Krings 2001) and most likely also due to the fact that in the think-aloud task participants could move backwards in the text.

Working memory capacity and French tasks were carried out after post-editing both texts, with the French vocabulary test being carried out last both in the eye-tracking and think-aloud tasks. While fatigue caused by the PE tasks could have had an influence on results of the French and working-memory-capacity tests, these additional tests were administered after the main PE task in an attempt to avoid fatigue effects on data reflecting cognitive effort in PE, the main variable in the study. In any way, the eye-tracking tasks were not sufficiently long (see section 4.1.1) for this ordering to be considered a cause for concern.

Additional background information about the participants, such as age and level of education, was collected by asking participants to fill in a post-task form. The form was completed after the working memory capacity task (in the case of the eye-tracking task) and it included a link to the French vocabulary test – see Appendix D. In an attempt to quantitatively control for effects of previous knowledge of the subject matter in the eye-tracking task, this form also asked participants to report if any previous knowledge they might have had on the topics of the texts had influenced their editing behaviour. This did not seem worth taking into account in the analysis in view of the low number of participants in the eye-tracking task reporting an influence of previous knowledge (3 out of 19).

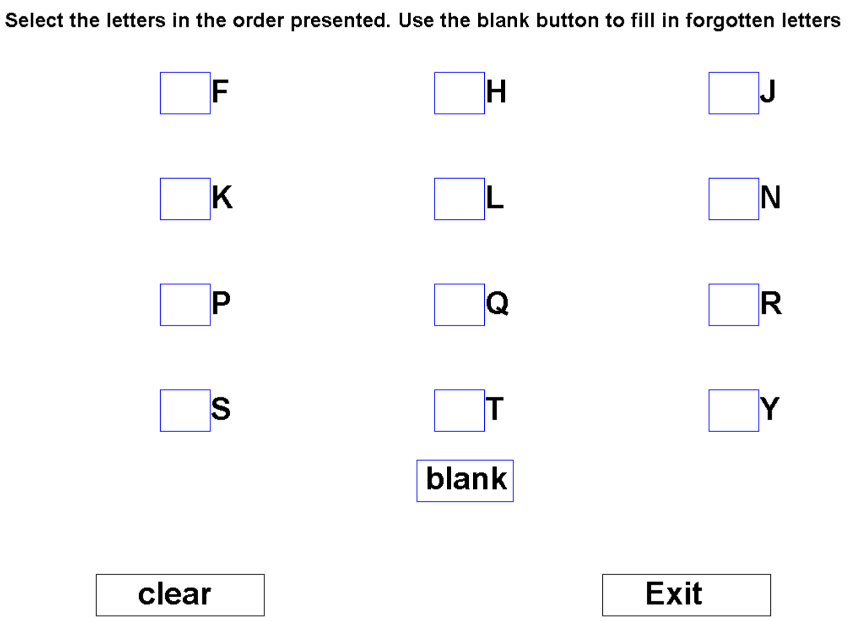
Details of the measurement of working memory capacity and French knowledge are provided below.

3.3.4.1. *Working Memory Capacity Task*

Working memory capacity was measured with the automated version of a reading span task (Unsworth, Heitz, Schrock and Engle 2005; Unsworth, Redick, Heitz, Broadway and Engle 2009), which is part of the ‘complex span’ group of tasks (Daneman and Carpenter 1980) normally used to measure working memory capacity. This task was used in a number of previous studies and has been found to produce reliable results (see Redick et al. 2012). It involves asking participants to recall sequences of between 3 and 7 letters that flash one at a time on screen in between sentences that need to be mentally processed. For each sentence, participants have to mark ‘true’ or ‘false’ as to whether or not the sentence is logical – e.g. ‘The prosecutor’s dish was lost because it was not based on fact’ (false) and ‘Throughout the entire ordeal, the hostages never appeared to lose hope’ (true).

The task has an initial training phase comprising three practice modes: just sentences, just letters, and both the sentences and letters. The time normally taken by each participant to process the sentences in the just-sentences practice mode is automatically taken into account and if, in the main task phase, the participant takes longer than expected (more than 2.5 standard deviations from the practice mean) to mark ‘true’ or ‘false’ for each sentence, this is considered a time error. Also, if a sentence that does not make logical sense is marked as ‘true’, or vice versa, this is considered a reading error.

The order in which sentences are presented is automatically randomised for each run of the test. At the end of a sequence of sentences, participants are asked to select the letters that flashed on screen in the correct order, as shown in Figure 7. If only parts of a sequence can be recalled, participants can use wildcards (a ‘blank button’) to replace any letters in the sequence that they cannot remember. Absolute and partial scores of working memory capacity are then obtained based on the number of letters remembered correctly, in the right position, in fully and partially remembered sequences, respectively. Results are usually considered valid only if time and reading errors are kept at a rate below 20%.[25]



Select the letters in the order presented. Use the blank button to fill in forgotten letters

Figure 7 - Interface of WMC task, where participants had to select a sequence of letters that had flashed on screen.

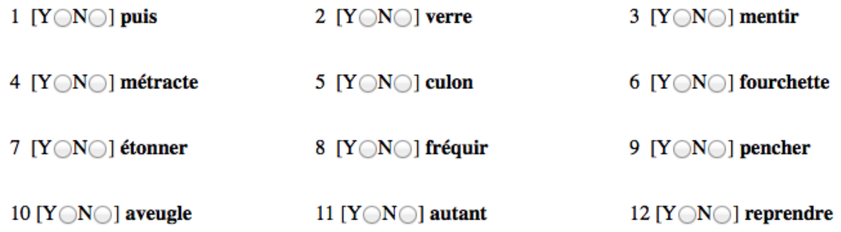
3.3.4.2. *French Vocabulary Test*

Participants’ knowledge of French was estimated with a yes/no vocabulary test[26] (Meara and Buxton 1987). The efficacy of using tests of this kind for the purpose of measuring foreign-language proficiency has been examined in a number of previous studies and this type of task is normally deemed robust ‘particularly for placement and

[25] The test was carried out twice for P04 as the number of errors committed by this participant exceeded the 20% limit. Using this participant’s second score was not seen as problematic, as both scores were low in comparison with those obtained by other participants in the study (see Table 4, p. 75).

[26] Available at http://www.lextutor.ca/tests/yes_no_fr/ (accessed November 2014).

diagnostic purposes' (Read 2007, 113). The task used here involves asking participants to mark 'yes' or 'no' as to whether or not they know the meaning of sixty words shown on screen. Forty of these words are true French words, while the other twenty are non-words that are nonetheless deemed to sound plausible to a non-native speaker of French – e.g. 'fréquir'. If positive responses are provided for non-words penalties are applied to the final score, which varies between 0 and 100. In the present study, participants were asked to take the first level of this test, where true words are sampled from a list of the one thousand most frequent words in French. Those who successfully passed level 1 of the test also took level 2, though scores in the second level were not used in the study for consistency across the participant sample. The interface used in the test is presented in Figure 8.



1 [Y○N○] **puis**	2 [Y○N○] **verre**	3 [Y○N○] **mentir**
4 [Y○N○] **métracte**	5 [Y○N○] **culon**	6 [Y○N○] **fourchette**
7 [Y○N○] **étonner**	8 [Y○N○] **fréquir**	9 [Y○N○] **pencher**
10 [Y○N○] **aveugle**	11 [Y○N○] **autant**	12 [Y○N○] **reprendre**

Figure 8 - Interface of French vocabulary test.

3.3.5. *Participants*

Ethical approval was obtained from the author's institution for the recruitment of human participants to take part in the research. Potential participants received an information sheet explaining any details that participating in the study would involve (see Appendix E). Those who agreed to take part were asked to provide formal consent by signing a form (see Appendix F). Nineteen participants (P01-19) were recruited for the eye-tracking task and an additional ten participants were recruited for the think-aloud task (P20-29), although one of these was left out of the analysis due to a difficulty in getting acquainted with the think-aloud condition.

All participants were native speakers of English and had previous experience with translation either as students or professionals. Their experience was mainly in translation and traditional revision, with only three participants (P13, P19 and P25) having professional experience in PE. As in a professional setting it seems common for PE to be carried out by translators (see Guerberof 2013), the low level of PE experience in the sample did not seem problematic. Also, similarly to that found by de Almeida

(2013), no obvious behavioural differences could be observed in the study between participants who had experience only in translation and those who had experience also in PE, so this distinction did not seem worth pursuing in the analysis.

All student participants (STD) were, at the time of the task, undertaking Modern Languages or Translation Studies higher education degrees[27] (at undergraduate or postgraduate level, referred to here as STD-U and STD-P, respectively). Professionals (P) were practising freelancers based mainly in the North East of England. The sample also included non-practising translators (NP) who had received formal higher-education training in translation. The NP group also includes P15 who, despite not having received formal training in translation, had passed a translation agency entry exam and was in the process of setting himself up as a professional.

Previous research in PE indicates that translators' attitude towards MT may influence their PE performance (de Almeida 2013). In view of this, participants were also asked to rate how inclined they were to using MT as a first draft in their daily translating practice (assuming this was possible/allowed), with 1 standing for 'not inclined at all' and 5 standing for 'I would certainly do it'. Naturally, estimating attitudes is not unproblematic. Since some participants were new to PE, it was deemed preferable to include this question in the questionnaire to be completed *after* the task had taken place as a way of avoiding having to provide extensive preliminary explanations which could influence their behaviour. As a consequence, it cannot be excluded that their experience in the task itself may have influenced their answers. Nevertheless, all participants were exposed to the same materials, under the same conditions, so any influence of the task itself on self-reported attitude scores would be expected to remain relatively constant.

The allocation of participants to one of the two tasks was carried out purposely in order to guarantee task samples with comparable profiles. This allowed for a more reliable contrast of eye-tracking and think-aloud data. In view of the relatively small number of participants recruited for the project and their diverse backgrounds, randomising participants' allocation to one of the tasks could lead to highly unequal task samples.[28]

[27] All of which including translation practice seminars.

[28] On one occasion, a circumstantial aspect also influenced the allocation of participants to one of the tasks: the fact that the type of glasses worn by P20 prevented a reliable collection of eye-tracking data, which influenced the inclusion of this participant in the group carrying out the think-aloud task, where eye tracking was not the main data-collection method relied upon.

Participants	Status	FR Vocab (0-100)	Experience (in years)	Attitude (1-5)	WMC (0-75)	Age
P01	STD-P	79	1.5	5	25	23
P02	STD-U	95	0	2	42	21
P03	STD-U	13	0	4	56	22
P04	NP	88	0	4	14	25
P05	NP	95	0	5	36	55
P06	NP	50	0.03	2	56	27
P07	STD-U	97	0.1	4	44	22
P08	STD-U	95	0	4	68	21
P09	STD-U	97	0	4	61	22
P10	STD-P	89	4	3	38	26
P11	STD-U	18	0	4	61	21
P12	P	60	13	2	32	40
P13	P	95	3	2	37	60
P14	NP	3	0	4	51	31
P15	NP	93	0	3	46	20
P16	STD-P	93	5	3	33	37
P17	NP	63	0.25	4	35	27
P18	P	97	4	2	50	30
P19	P	18	10	5	39	69
Mean ET-P task (N = 19)	-	*70.4*	*2.1*	*3.4*	*43.3*	*31.5*
P20	P	95	13	2	-	38
P21	NP	75	0	3	-	25
P22	STD-U	90	0	3	-	21
P23	STD-U	10	0	5	-	23
P24	P	97	0.6	3	-	25
P25	P	98	28	2	-	67
P26	P	93	10	4	-	46
P27	STD-P	98	0	5	-	66
P28	STD-P	78	0.1	3	-	26
Mean TA task (N = 9)	-	*81.5*	*5.7*	*3.1*	-	*37.4*

Table 4 - Participants’ profile.

Participants’ age, level of professional experience, French vocabulary test result and scores reflecting attitude towards MT are presented in Table 4 for both tasks, together with the absolute working-memory-capacity (WMC) scores of those taking part in the eye-tracking task. Relatively small differences can be observed between the two participant samples, with a higher level of professional experience and higher level of French for think-aloud participants, who were also older on average. These differences, however, were not found to be statistically significant based on non-parametric Mann-Whitney-Wilcoxon tests, which are robust for non-normal data as well as outliers.[29]

3.3.6. *Post-Edited Quality*

3.3.6.1. *Assessment Design*

The MT output as well as all post-edited versions produced in the study were evaluated by human assessors. The evaluation tool developed within TAUS's Dynamic Quality Evaluation Framework was used to assess the translations on 1-4 (low-high) scales of fluency and adequacy. The scales used are those described in section 2.3.2. Fluency 'captures to what extent the translation is well formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker', while adequacy measures 'how much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation' (Linguistic Data Consortium 2005, in TAUS 2013a).

Assessors were exposed to pairs comprising a source sentence and corresponding translation. Raw-MT and post-edited versions were randomly scrambled whilst following the internal order of each text. The sentences were divided into six batches of approximately 250 source-and-translation pairs each. Two 'mismatched' sentence pairs that had little to no correspondence between French and English were deliberately included in each batch as a way of estimating the reliability of the assessment, i.e. by checking to see if these pairs received low adequacy scores. An example of such pairs is provided below.

ST: *Des études sur la vitamine E sont contradictoires, selon la SCC.* 'Studies on vitamin E are contradictory, according to the SCC.'

Target: *While some studies noted a decreased risk of prostate cancer, another rather noted an increase in vitamin E levels.*

Five self-described professional translators were initially hired through oDesk[30] to carry out the assessment. The oDesk website allows the advertisement of jobs to which professional freelancers can apply using profiles they create on the website itself.

[29] French Vocabulary: W = 65.5, p = 0.33; experience: W = 75.5, p = 0.62; attitude towards MT: W = 102.5, p = 0.40; age: W = 67.5, p = 0.39.

[30] This website is currently called 'UpWork' – see https://www.upwork.com/ (Accessed 29 May 2015), formerly https://www.odesk.com/info/uk/welcome/ (Accessed 23 May 2014).

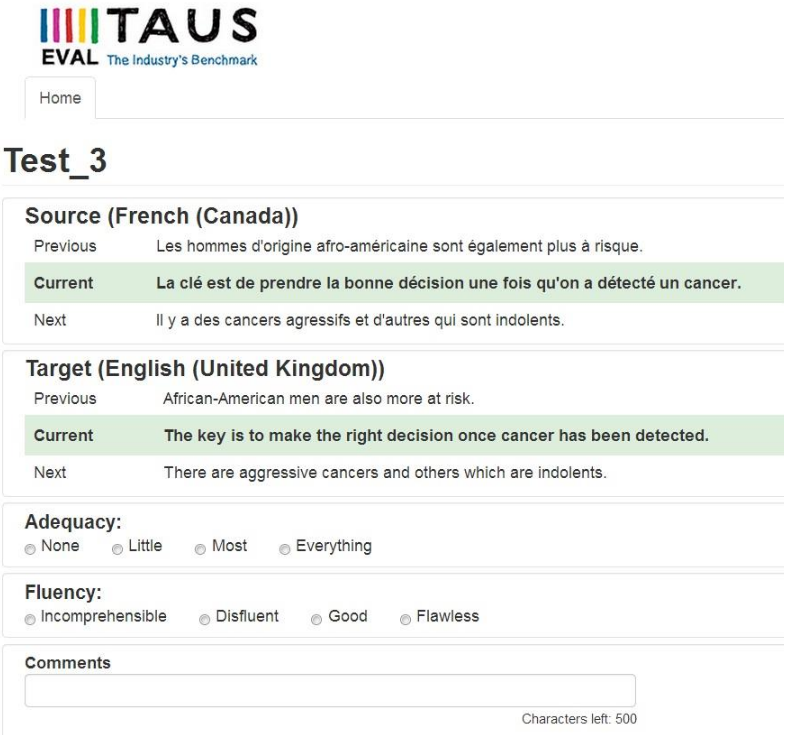
Out of the five judges initially hired, three were kept in the study. The assessment carried out by one judge was deemed unreliable, as out of twelve mismatched sentence pairs included in the data, three were rated with 4 out of 4 on adequacy. As for the second judge not kept in the analysis, in quite a few instances comments were left in allusion to linguistic problems in the translations, but in some cases the errors identified did not reflect the score provided, which was often 4 out of 4 on fluency.

The three judges (J01-03) retained in the analysis had previous professional experience in French-English translation. One of them also had experience in rating the linguistic accuracy of web search hits. They were all native speakers of English and they had at least an undergraduate degree in French, with J02 also having a postgraduate degree in French-English translation. They had previously taken French-English translation skills tests provided by oDesk[31] and they were also asked to take the same French vocabulary test given to post-editors. Judges' French vocabulary and translation skills test scores are presented in Table 5 and TAUS's translation assessment interface is presented in Figure 9.

	FR Vocabulary (0-100)	FR-EN Translation Skills (0-5)
J01	93	4.6
J02	100	4.5
J03	89	4.9

Table 5 - Results of French vocabulary and translation skills tests for each judge.

[31] These are multiple-choice tests that can be taken by those with a profile on the website as a way of highlighting their skills to potential clients. The French-to-English translation test comprises 40 questions and covers aspects such as the use of prepositions, indirect speech and time expressions.



Test_3

Source (French (Canada))

Previous	Les hommes d'origine afro-américaine sont également plus à risque.
Current	**La clé est de prendre la bonne décision une fois qu'on a détecté un cancer.**
Next	Il y a des cancers agressifs et d'autres qui sont indolents.

Target (English (United Kingdom))

Previous	African-American men are also more at risk.
Current	**The key is to make the right decision once cancer has been detected.**
Next	There are aggressive cancers and others which are indolents.

Adequacy:
○ None ○ Little ○ Most ○ Everything

Fluency:
○ Incomprehensible ○ Disfluent ○ Good ○ Flawless

Comments

Characters left: 500

Figure 9 - Interface of TAUS Dynamic Quality Framework tool, used for fluency and adequacy assessment.

The judges were asked to do two batches per day, completing the entire task over a fixed period of three days. The order of presentation of assessment batches was alternated between judges over task days – i.e. the two batches assessed by a given judge on day one were assessed by the following judge on day two and so forth. An unlimited break could be taken between batches, but judges were asked to do each batch in a single sitting. No information on the origin of the translations was provided – i.e. if they were raw MT outputs or post-edited versions. Before starting the assessment, judges were asked to read a briefing document (see Appendix C) including scoring examples and an explanation of the fluency and adequacy scales adopted. They were asked to assess fluency first, without looking at the source sentence, and only then compare source and target sentences to evaluate adequacy. They were also asked to read, prior to the evaluation, the texts with subject-matter information post-editors had access to as well as reference English translations. They were instructed not to consult the reference translations during the task, however, as it would be difficult to control how much each judge would rely on it, which could unbalance the assessment.

3.3.6.2. *Inter- and Intra-Rater Reliability*

Assessing translation quality is of course a highly subjective task, so fuzziness is expected both between assessors and within a single assessor’s evaluation. In view of this, reporting the reliability of the assessment seemed desirable. For this purpose, inter- and intra-rater reliability was measured with Krippendorff's alpha (α) (Krippendorff 2011), a reliability test that can be used for ordinal scales. Traditional methods for measuring intra- and inter-rater reliability for samples of more than three assessors, such as Fleiss's κ, only take perfect agreement into account, not differentiating between disagreements with different levels of severity – e.g. a disagreement between 1 and 4 (more severe), and a disagreement between 3 and 4 (less severe). Krippendorff's α is able to account for these differences and accommodate partial agreement, so it was the measure chosen here.

A small number (39) of duplicate source-and-translation (French-English) sentence pairs were included in the data to allow for the measurement of intra-rater reliability, i.e. by checking to see if the same score was provided by the same participant on both occasions where these pairs were seen.

	Inter-Rater Reliability		Intra-Rater Reliability	
	Fluency	Adequacy	Fluency	Adequacy
α	0.171	0.519	0.762	0.636

Table 6 - Intra- and inter-reliability for post-edited fluency and adequacy.

Inter- and intra-rater reliability results are presented in Table 6.[32] A Krippendorff's α of 0 represents chance agreement, while an α of 1 represents perfect agreement. Inter-rater reliability results for adequacy (α = 0.519) were considerably higher than those obtained for fluency (α = 0.171). This could be indicating that grasping a notion of adequacy may be more straightforward, since this concept is judged based on the proportion of shared meaning between source and target. Assessing the linguistic quality of the translations, on the other hand, seems to involve higher levels of subjectivity, potentially requiring more extensive training if higher reliability is to be achieved. Inter-rater reliability for fluency is nevertheless above chance. The reliability achieved for adequacy could be considered moderate.

As it would be expected, intra-reliability results are substantially higher than inter-reliability ones, which can be observed both for fluency (α = 0.762) and adequacy (α = 0.636). This indicates that judges were consistent in their own scoring.

[32] These refer just to data obtained in the eye-tracking task, which is the primary source of information for addressing RQ5 (i.e. linked between PE effort and post-edited quality).

Regarding the inter-rater result obtained for fluency, it seems indeed that low agreement between translation-quality evaluators is not uncommon in previous research. Carl et al. (2011), for example, observed agreement rates that were only marginally above chance. Discrepancies of this kind are not regarded as problematic in the present thesis as fluency and adequacy results are considered here based either on modal or average scores (i.e. the most frequent score for a given sentence or the average of all three judges, respectively) or on mixed-effects models, where variation between judges is controlled for (see section 3.4.5.2).

It is also worth mentioning that while all judges were native speakers of English, the final selection includes a national of Canada (J01), of the United States (J02) and of the United Kingdom (UK) (J03). This was not anticipated as a problematic scenario, though admittedly a UK-only sample would have arguably been a better approach in obtaining consistency with the sample of post-editors, who were mostly from the UK (P19, a participant from the United States, being the only exception). In any way, this was not deemed a cause for concern as after subsequent checks it was noted that similar results on post-edited quality would be obtained if the analysis was based on the UK judge alone.

3.4. Data Coding and Processing

In this section, the steps involved in coding and processing the data are described. The process of extracting textual features from the ST and the MT output is explained in section 3.4.1. Section 3.4.2 describes how eye-tracking data was filtered and processed. Section 3.4.3 describes the specific measures exploited as indicators of cognitive effort in the eye-tracking task. Preliminary steps involved in processing think-aloud data are outlined in section 3.4.4. Finally, section 3.4.5 describes the statistical analysis approach adopted in the study.

3.4.1. *Linguistic and Psycholinguistic Predictors of Cognitive Effort*

3.4.1.1. *Predictors based on the ST*

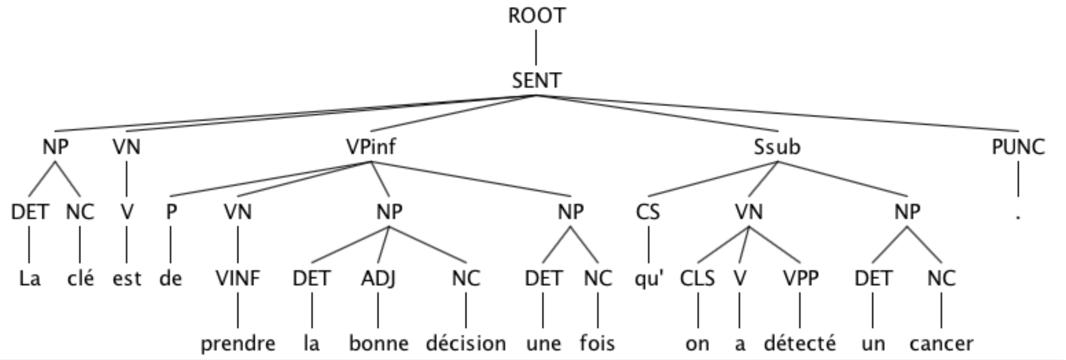
In line with previous research, ST characteristics investigated in the context of RQ1 (i.e. links between textual characteristics and cognitive effort) consist of linguistic features that can be extracted automatically from texts. In the present study, the use of relatively simple features was prioritised and it also seemed desirable to test features that occur frequently in the language (i.e. features that are not expected to be related to a specific type of text or context), as this accounts for simplicity in the statistical analysis. Variables that seemed to meet these criteria include features based on part-of-speech categories, lexical frequency (i.e. percentage of words belonging to a list of the one thousand most frequent words in French – see section 3.3.2), and relatively basic features such as lexical density (i.e. a ratio of content words over all words) and type-token ratio (i.e. a ratio of individual tokens occurring in the text, without repetition, over the number of all tokens, including repetitions).

Word frequency has been used as a potential indicator of translating difficulty in previous research (Hvelplund 2011) and is also known to influence eye-movement behaviour in reading (Rayner 1998), hence its use here. Type-token ratio could predict effects posed by the repetition of words in the sentences, while lexical density may index the impact exerted by the amount of information conveyed by the ST, with higher density being expected to require more mental processing capacity, hence more cognitive effort. Features based on part-of-speech categories have been largely used in previous research aimed at predicting PE effort (e.g. Aziz, Koponen and Specia 2014; Green, Heer and Manning 2013), so their use here also seems largely justified.

All of the features described above were extracted from the STs at a sentence level. Type-token ratio, lexical density and frequency were obtained from a vocabulary profiler available within the *Compleat Lexical Tutor* online tool.[33] The frequency of different part-of-speech categories was obtained by parsing the source sentences with Stanford French Parser (Green et al. 2011), a tool that tokenises[34] and syntactically parses French text, classifying individual words as to their part-of-speech category as well as groups of words as to their phrasal categories.

[33] Available at http://www.lextutor.ca/vp/fr/ (accessed December 2013).

[34] Tokenisation is a process where the text is broken into single individual particles, such as breaking the string *l'état* into *l'* + *état*.



SENT: sentence; NP: noun phrase; DET: determiner; NC: noun; VN: verbal nucleus; V: verb; VPinf: infinitive clauses; ADJ: adjective; CS: subordinating conjunction; CLS: subject clitic pronoun; VPP: participle; PUNC: punctuation.

Figure 10 – Parse tree of sentence 10, with dependency relations and part-of-speech labels.

The parsed output of sentence 10 (text A) is presented above in Figure 10, where acronyms in the parse tree are phrase and part-of-speech labels. After parsing texts A and B, the frequency of different phrases and part-of-speech categories in each sentence was examined. Certain features were not deemed frequent enough in the sample for their use to be justified in the study, which was observed, for example, for multiword expressions. Vieira (2014) tested each feature yielded by the parser individually, but here it seemed desirable to merge infrequent features into overarching categories as a way of simplifying the analysis. A list of the part-of-speech and phrase features tested in the study is presented below.

N (NC+N+NPP) – nouns, including proper nouns and nouns included in multiword expressions.

V (V+VINF+VPP) – verbs, including participles and infinitives.

ADJ – adjectives.

ADV – adverbs.

P – prepositions.

VP (VPinf + VPpart) – verb phrases, including infinitive and non-finite clauses.

NP – noun phrases.

AP – adjectival phrases.

PP – prepositional phrases.

As the incidence of these features in each sentence would be expected to vary with sentence length alone, the features were normalised by the number of tokens in the source sentences. Raw counts of these features per sentence are provided in Appendix G.

3.4.1.2. *Predictors based on the MT Output*

Meteor is the primary MT-output feature tested in the study. Since the translations used as reference were produced by human professional translators, Meteor is expected to give a good estimate of the levels of quality of the MT output, shown in previous research to have a complex relationship with the amount of cognitive processes that take place in PE (Krings 2001, 539).

The MT output was also exploited to supply linguistic features obtained with version 3.0 of the Coh-Metrix automatic text analysis tool[35] (Graesser et al. 2004; Graesser and McNamara 2011). Even though linguistic features based only on the ST are analysed in the majority of previous research, both practical and theoretical reasons have motivated the decision of also using the raw MT output here for this purpose. First, to the knowledge of the present author, tools capable of providing advanced indices such as those available within the Coh-Metrix tool are not currently available for French, the source language adopted for the study. Second, though characteristics of the ST may be mirrored in the MT output, it also seemed interesting to test linguistic features based directly on the machine-translated text. In this context, these measures could reflect the complexity of the MT output which, after the emerging revised TT, is the textual component participants primarily focus on (see section 2.1). A list of the Coh-Metrix indices selected for analysis is provided below. While Coh-Metrix provides a large number of different features, only those that could be applied to separate sentences, as opposed to entire texts, were considered.

WRDPOLc – Average polysemy for content words, based on WordNet (Miller 1995).[36]

WRDHYPnv – Mean hypernymy of nouns and verbs, measured as the number of hierarchical levels in associations maintained with other words with broader meanings in WordNet, i.e. how specific words are.

WRDAOAc – Age of acquisition of content words, i.e. the age from which native speakers of English are normally expected to have the word in their vocabulary.

[35] Available at http://cohmetrix.memphis.edu/cohmetrixpr/index.html (accessed December 2013).

[36] See http://wordnet.princeton.edu/ (Accessed July 2015).

WRDFAMc – Familiarity score of content words, i.e. how familiar the words sound to an adult.

WRDCNCc – Index of word concreteness.

WRDIMGc – Index of word imagability, i.e. how easy it is to construct a mental image of the word.

WRDMEAc – Word meaningfulness ratings (Toglia and Battig 1978).

3.4.1.3. *Principal Components Analysis*

To avoid over-fitting and multicollinearity problems in the statistical analysis – i.e. too many or inter-correlated predictors – Coh-Metrix indices plus ST lexical frequency, lexical density and type-token-ratio were treated with principal components analysis[37] prior to entering the statistical models used in the study. Principal component analysis is a statistical technique that takes a set of *n* variables as input and aims to provide a small number of orthogonal (i.e. uncorrelated) principal components (PCs) as output. Seven PCs, accounting for 95% of the variance in the original features altogether and at least 5% each (Baayen 2008), were selected for further analysis. These PCs are provided in Appendix G. Syntactic features obtained by parsing the texts were not included in the principal component analysis as this made it easier to specify which part-of-speech categories could be driving any effects. This seemed of interest in view of the relative simplicity of part-of-speech features and their frequent use in previous research.

3.4.2. *Gathering Eye-Tracking Data*

The use of eye tracking requires that a number of steps be taken to increase confidence that the data indeed reflects cognitive processes. In particular, the equipment needs to be calibrated for each participant and physical conditions in the eye-tracking room need to be appropriate, e.g. light sources being kept constant (see Hvelplund 2011; O’Brien 2010). Fixed chairs may also be used to prevent participants from making abrupt movements (Hvelplund 2011, 103). All these measures were taken into account in the present study. The distance between the eye tracker and the screen as well as other set-up details were established based on Tobii’s configuration guide[38] (see section 3.4.2.1, below). Calibrating the eye tracker involves asking participants to look at a moving object on screen and checking the precision of the gaze based on five fixed

[37] Using the `prcomp` R function with options `scale` and `center`.

[38] Available at http://www.tobii.com/en/eye-tracking-research/global/library/manuals/ (accessed November 2014).

points, one in the centre and one in each corner of the monitor. The calibration process required between 10 and 30 minutes for each participant.

3.4.2.1. *Eye Tracker and Fixation Filter Settings*

The eye-tracking equipment used in the study was a remote non-invasive Tobii X120 eye tracker, which collects gaze samples at a rate of 120Hz. This means that the equipment looks for participants' eyes and generates a raw gaze data point at approximately each 8.3 milliseconds. Prior to any analysis, the raw data needs to go through a filtering process in which fixations are distinguished from saccades (see section 2.2.3.1) or from other data points that simply constitute noise or points in time when participants' eyes are not found – e.g. due to looking away from the screen to use the keyboard. Tobii I-VT was the filter used in the present study (see Tobii Technology 2012). This filter classifies raw data points into fixations based on how close together these points stay on screen and for how long. In line with previous research in translation (Hvelplund 2011) and with settings adopted in the EYE-to-IT project,[39] in the present study the filter was set to discard individual fixations with durations below 100 milliseconds. All other fixation filter settings were kept at default values.

3.4.2.2. *Processing and Annotation of Gaze Data*

Tobii Studio was used to record the screen as well as participants' verbalisations (in the think-aloud task), which were all available from .avi files that also included superimposed gaze information (where participants looked on screen). Tobii Studio automatically produces a number of different eye-tracking metrics that can be further exploited for analysis, such as fixation count or time to first fixation. When the analysis is based on a video, these metrics need to correspond to selected scenes within the video whilst also pertaining to specific areas of the screen, normally referred to as areas of interest. In the eye-tracking task, the process of post-editing each separate sentence was marked as a scene in the task video by annotating the recordings with labels corresponding to the start and end of the PE process of each sentence. The horizontal ST-TT rectangle on PET's interface (see Figure 5, p. 68) was selected as an area of interest. The ST and MT output/TT areas were individually established as two additional areas of interest, which allowed the collection of gaze data pertaining specifically to the ST or the MT output/emerging TT. Eye-tracking metrics were then automatically obtained with Tobii Studio for each area of interest, within each scene.

[39] See http://cordis.europa.eu/fp7/ict/fet-open/portfolio-eyetoit_en.html (accessed November 2014).

When remote eye trackers are used, the estimated landing of the gaze on screen may suffer slight drifts (see e.g. Sjørup 2011). To account for this, enough space was left between the ST and the MT output on PET’s interface to guarantee a reliable collection of data pertaining specifically to one of these elements.

3.4.2.3. *Quality of Eye-Tracking Data*

Despite taking a number of measures to prevent low-quality data, it is generally the case that some of the data produced in eye-tracking studies needs to be disregarded due to failing to reach certain standards of quality. In the present study, five data points (10% of P19’s sentences plus 2% of P14’s) were discarded from the outset due to accidental errors in task execution that could increase fixation time, such as transposing the order of sentences and clicking on the wrong buttons. Furthermore, the data was checked for any measurement noise that could mislead the analysis, with average fixation duration and ratios of fixation time over total editing time being exploited as data quality parameters. These parameters were examined for each ST-MT sentence pair, with data points that did not conform to either measure being excluded.

The use of average fixation duration as a data quality measure is based on findings from reading research, where 225 milliseconds (Rayner 1998) is the average duration normally observed, with a minimum threshold of 200 average milliseconds being generally adopted as a quality parameter in previous research in translation and PE (Hvelplund 2011; O’Brien 2011). This minimum threshold was also followed in this thesis.

The use of fixation time over task time to measure eye-tracking data quality consists of checking to see if enough data is available for analysis. In line with the procedure followed by Hvelplund (2011), abnormally low fixation time was identified in the present study based on cases that considerably deviated from the sample mean.

Of the 774 data points deriving from the eye-tracking task (i.e. 19 participants x 41 sentences, minus five exclusions due to task execution errors), 59 (8%) were discarded due to being below the minimum average fixation duration of 200 milliseconds. A further 7 data points (0.9%) were excluded because the ratio of fixation time on screen over total editing time was more than 2.5 standard deviations (0.14) below the mean of the sample (0.64). The time spent fixating on the text was lower than 29% of total editing time for these data points, which indeed seemed to denote a case of noise as it would be unlikely for participants to be looking away from the screen 71% of the time.

3.4.3. *Estimating Cognitive Effort in the Eye-Tracking Task*

The rationale for the use of eye movements and subjective ratings as measures of cognitive effort has been discussed in section 2.2.3. A number of such measures is available and specific ones had to be selected. Here, again, it seemed desirable to triangulate sources.

With regard to eye tracking, average fixation duration and fixation count (for each ST-MT sentence pair) were the metrics selected. The reasons for these specific choices are two-fold. First, these measures are largely used for similar purposes in related work (e.g. Doherty, O'Brien and Carl 2010; O'Brien 2011). Second, previous research suggests that the average duration of fixations tends to produce slightly different results in relation to other eye-tracking measures, potentially reflecting a different facet of cognitive processing (Doherty, O'Brien and Carl 2010; van Gog, Kester, Nievelstein, Giesbers and Paas 2009). This seemed to be particularly interesting for the purpose of triangulating data (Creswell 2009), as this measure could potentially provide a slightly different angle of observation.[40]

Based on the assumption that individuals are able to introspect and report in terms of numerical values the levels of cognitive effort they expend (Gopher and Braune 1984, in Paas 1992), a subjective scale of mental effort from the field of educational psychology was adopted as another source of data reflecting cognitive effort. This scale varies from 1 'very, very low mental effort' to 9 'very, very high mental effort' (Paas 1992). For details of previous experiments where this same scale has been used see Roodenrys, Agostinho, Roodenrys and Chandler (2012) and Tabbers, Martens and van Merriënboer (2004). As previously mentioned, participants were automatically prompted to provide these scores by clicking on levels of the scale on screen after editing each sentence (in the eye-tracking task only). Internal numerical points in the scale were not labelled in the present study. Verbal labels do not make a significant difference in results according to Paas (1992) and their absence was deemed desirable here as it reduced the amount of content to be read on screen in between sentences, avoiding a high degree of distraction from the main task. The variable consisting of these subjective ratings is henceforth referred to as 'subjective cognitive effort'.

[40] Indeed, only a weak correlation was observed in the present study between fixation count and average fixation duration (see section 4.1.2).

3.4.4. *Processing and Coding TAPs*

The procedure adopted for analysing the TAPs follows the traditional steps of transcription, coding and analysis (Sun 2011, 943), with segmentation also being carried out prior to coding (Krings 2001, 309-310). These stages are described below in sections 3.4.4.1 and 3.4.4.2.

3.4.4.1. *TAPs Segmentation*

Verbal data resulting from the think-aloud task was transcribed in NVivo,[41] a tool for qualitative data analysis. After an initial transcription of the data, key logs produced with Translog-II as well as the task videos themselves were exploited to extend the transcription with information on any edits performed by participants. This step seemed necessary as on a number of occasions the edits implemented were not directly verbalised, which could potentially reflect automatic processes not available for reporting (see Jones 2011). Information of this kind seemed relevant in the context of PE, where correcting the MT output may involve carrying out brief local edits due to linguistically superficial issues.

After transcription, criteria that could be used to segment the data into coding units were operationalised. Segmentation rules put forth by Krings (2001, 309-310) were taken into account for this purpose. Though coding units in Krings's study serve a slightly different purpose than the one they serve here, these rules seemed particularly suitable as they were also designed for PE.

Pauses are the main segmenting criterion taken into account by Krings. Following previous psycholinguistic research, Krings establishes pauses of one second or more as a unit delimiter in TAPs. Further to pauses, he suggests a number of other delimiters, including attentional shifts between source and target languages, different editing alternatives/solutions, shifts between the object language (i.e. the ST, MT or edited/emerging TT) and participants' own commentaries, and shifts to or from physical writing processes. In the present study, physical writing actions were only treated as separate units when they were not orally reported.

Krings also establishes that different propositions should form separate units. Krings defines propositions as 'elementary statements consisting of a predicate and several arguments' (2001, 238). As a segmentation criterion, this is also considered by Krings in the context of a final fusion rule whereby the pause rule is disregarded if

[41] See http://www.qsrinternational.com/products_nvivo.aspx (accessed 15 April 2015).

verbalisation units are deemed to be homogenously connected, forming a single proposition. This principle was also followed here, though with most connective propositions (i.e. groups of simple propositions connected by conjunctions – see footnote 2) being broken into separate coding units, which seemed desirable as a way of allowing for a more precise coding of the data. The only exception in this respect was when a clause did not seem to constitute a logical statement as a standalone unit, e.g. in the case of clauses linked together by the conjunctions ‘if’ and ‘that’, where keeping both main and subordinate clauses in the same unit seemed necessary.

An extract of P20’s post-editing of text B is presented below in Ex.1, where each line corresponds to one coding unit (U) according to the segmenting criteria outlined above.

(Ex.1) U315 I wonder I think *republican* will have a capital *R*
U316 because it’s a political party
U317 uh not ~~*defined*~~ *a strategy*
U318 maybe ***drawn up***
U319 or *set out*
U320 anyway
U321 *drawn a up a strategy to ensure that the presidency*
U322 *a mandate*
U323 no
U324 *one*
U325 *a single mandate*
U326 I don’t know if you’d really say that
U327 *a* ***single term*** [~~*mandate*~~]

KEY:[42]
bold: insertions; ~~strikethrough~~: deletions; *italics*: ST, MT, or edited/emerging TT; [ ]: physical writing processes not spoken out loud or comments added by the researcher; underline: participant’s emphasis

3.4.4.2. *Coding*

The coding strategy adopted in the study could be regarded as a mixture of both inductive and deductive approaches. While coding was directly motivated by RQ3 (i.e. linguistic aspects of the task participants attended to), room was also left for the coding process to be influenced by the data, with unexpected cases being accommodated accordingly.

The present author preliminarily coded all units. An external coder was then involved in the analysis, which served the purpose of fine-tuning the coding scheme and

[42] This same key applies to all examples of TAPs provided in this thesis

measuring inter-coder reliability. The external coder was a native speaker of English and had a PhD in French Studies. At an initial stage, the external coder was provided with a random sample of TAP sequences consisting of 100 coding units together with a proposed set of coding instructions comprising a description of each category and general criteria. Cohen's Kappa (a measure of inter-coder reliability[43]) for this initial sample was 0.56. After discussing cases of disagreement and making slight adjustments in the coding scheme and criteria, a new random sample of 100 units not previously seen by the external coder was used for a further check. This time a Kappa of 0.63 was achieved, which is within the band of the score described by Landis and Koch (1977, 165) as 'substantial' agreement, being therefore regarded here as satisfactory.

The coding scheme arrived at is divided into two major groups: specific foci and non-specific foci. The term 'focus' is used here to refer to different aspects of the PE activity participants consciously or automatically attend to, as indicated by their verbalisations or modifications implemented in the text. Specific foci units are those in which a specific linguistic aspect of the activity was evident. Non-specific foci are those where specific information on the linguistic nature of the issue dealt with was not explicated by participants.

It was established that each unit should be coded with a single category and that specific foci categories were supposed to be chosen over non-specific ones provided this was supported by evidence in the data. Individual coding categories are presented below together with examples.

Specific Foci

Lexis

Units (U) involving notions of lexical meaning, including collocations and the meaning of fixed structures/expressions. Units coded with this category generally involved content words:

(Ex.2)	U560	it's kind of a euphemism that has not been translated suitably
	U1478	I'm trying to think what the word is in English for that

[43] Cohen's Kappa normally varies between 0 and 1, with 0 standing for chance agreement and 1 standing for perfect agreement.

Grammar/Syntax

Units relating to grammar or syntax, generally involving function words – henceforth ‘Grammar’. Units in this category do not necessarily involve the correction of errors, but any changes in the text that could be regarded as relating essentially to grammar or syntax, such as those involving passive/active voice, plural, verb tense, etc.:

(Ex.3) U395 ***These*** [~~*This*~~] *new arrangements will influence*
U24 ah, word order is wrong [~~*States*~~]

Discourse

Units involving punctuation, consistency, the relationship between sentences, and other aspects pertaining to the overall coherence of the text:

(Ex.4) U535 I wonder how those two sentences were put together
U22 too many commas?

Style

Units relating to the flow/style of the text. Units under this category involved issues not deemed to directly fit the Grammar category, including the length of sentences, the amount of words used, repeated words close to each other, the inclusion of words that were not in the French original as a way of improving the flow of the English text, and shifts in word order which did not involve grammatical modifications (e.g. ‘therefore recommend’ instead of ‘recommend therefore’):

(Ex.5) U491 I think ~~*such a*~~ *requirement* is not necessary
U2066 that sounds too wordy

Translation Context/World Knowledge

Units pertaining to the specific context of the translation, involving aspects such as readership, genre-specific issues, source- and target-culture-specific issues, real-world use of the text, knowledge of the subject matter, intertextuality, etc. – henceforth ‘Knowledge’:

(Ex.6)	U497	I’d probably check whether *constitutionality* is used in America
	U15	*prostate cancer* is better for the headline

Orthography/Capitalisation/Typography

Units involving surface-level aspects of the text such as spelling, capitalisation and the use of spaces – henceforth ‘Orthography’:

(Ex.7)	U446	I think it’s only one space after the full stop
	U146	I presume African and American is they’re both capitalised [sic]

Non-Specific Foci

Non-Specific Reading/Evaluation

Three distinct reading modes were evident from the TAPs: the initial reading necessary to put text segments into working memory, reading that was connected to a specific task focus or problem, or final reading events where participants were simply checking if interventions in the text had worked or if any problems still remained. While the second mode was normally coded with one of the specific foci outlined above, the first and last modes were frequently coded with this category (henceforth ‘Reading’) as their motivation was often a neutral one, with a connection between the act of reading and one of the specific foci described above not being readily evident. The same applies to positive/negative evaluations whose motivation was not specified:

(Ex.8)	U81	and I think the last sentence is OK
	U4957	*because the cancer is not aggressive and does not threaten their lives* [initial reading]

Procedure

Units involving procedural aspects such as statements in which participants declared what they were going to do next:

(Ex.9) U1348 I'm going to read this in French
U2035 OK I'll come back to that

Units that were not deemed to fit any of the categories described above were coded with an 'Undefined' category.

Overall Coding Principles and Guidelines

In the interests of coding consistency, a number of guidelines were established to account for exceptional cases observed in the data. For example, when the same coding unit involved more than one textual modification, the category that was evident earlier in the unit was deemed to prevail. Similarly, when a single modification could be deemed to involve two categories at the same time (e.g. a change in word order together with a lexical substitution), the category pertaining to the earliest corresponding textual element in the MT output was considered to prevail. It was also established that the text being replaced should be used as a guiding principle, so if, for example, a pronoun was replaced with a content word, the unit in question was coded with Grammar as opposed to Lexis. If the inserted content word was later substituted by another, then Lexis was the category used.

Overall, as in most analyses based on TAPs, it cannot be excluded that a certain degree of fuzziness remains in the data, which may apply to both coding and segmentation. This, however, is alleviated by the fact that the dataset used here is relatively large (over 5,000 coding units, in total), so any fuzzy cases that may still remain would not be expected to significantly influence results. In addition, the inclusion of an external coder in the analysis, a procedure not often carried out in previous research (see e.g. Li 2004), is expected to have drastically reduced fuzziness in coding.

3.4.5. *Statistical Analysis*

3.4.5.1. *Multiple Regression*

The majority of research questions addressed in the present study entail comparisons involving multiple variables. Settings of this kind are more robustly handled in the context of multiple regression. This is a statistical technique that allows looking into potential correlations between a number of 'predictors', on the one hand, and a given 'response' (or 'outcome') variable, on the other hand – e.g. the impact of various ST and MT characteristics (the predictors) on a given measure of cognitive effort (the response variable). Regression analyses eliminate the need for running separate two-way correlation tests, allowing all potential factors that may influence a given response variable to be considered concomitantly. This means that any potential inter-relationships that may exist amongst predictor variables can be accounted for in the context of regression – e.g. the fact that ST characteristics themselves might influence MT quality, a situation where multiple regression allows estimating the individual impact that each of these elements might have on cognitive effort.

Regression was the main statistical technique adopted in this thesis. In practical terms, this means that most relationships investigated in the study were analysed by means of regressing a number of potential predictors on a given response variable. Cognitive effort is the overarching response variable in RQ1 and RQ2. The occurrence probability of different TAP categories is RQ4's response variable, observed as a function of cognitive effort. RQ5 has post-edited quality as response variable, being observed as a function of cognitive and other types of effort. RQ6 concerns the potential of TAPs for functioning as an index of cognitive effort, having the frequency of TAP units as response variable and different measures of cognitive effort as predictors. In brief, all research questions in the study are addressed primarily based on regression apart from RQ3, which does not involve correlational comparisons between variables, but rather the nature of different TAP categories and their frequency. A summary of the statistical comparisons carried out in RQ1-2 and RQ4-6 is presented below in Table 7.

	Overarching predictors	Overarching outcome
RQ1 & RQ2	ST, MT and participants' characteristics	cognitive effort
RQ4	cognitive effort	probability of diff. TAP categories
RQ5	cognitive and other types of effort	PE quality (fluency and adequacy)
RQ6	cognitive effort	TAP unit frequency

Table 7 - Summary of statistical tests carried out in the study.

Naturally, a number of other variables had to be controlled in the investigation, being therefore also included as predictors in the regression models used. These often comprised, for example, source-sentence length, which alone is expected to influence fixation count or the number of TAP units produced per sentence. In addition to being the main object of analysis in RQ2, participants' individual characteristics are also treated as control variables in the context of other questions addressed in the study.

Further to allowing multiple variables to be statistically controlled for, multiple regression also makes it possible to check for interactions between variables. Interactions take place when the effect of a given predictor on the outcome variable is moderated by other predictors. For example, if the levels of effort expended by non-professional post-editors were significantly higher when they edited long MT sentences, this could constitute an interaction between (1) extent of professional experience in PE and (2) MT sentence length. Checking for interactions of this kind significantly increases the complexity of regression models and if many of them are examined large amounts of data are required. In view of this, interactions were only checked in the present study when they were justified by previous research or when it seemed logical to check for effects of this kind (see next chapter).

In multiple regression models, it is important to avoid collinearity, i.e. correlated predictors, which can cause numerical imbalance in the model and mislead the analysis. In the present study, collinearity was checked based on the κ condition number, which indicates collinearity levels for several variables at once (i.e. an alternative to carrying out separate two-way correlation tests). Baayen (2008, 182) suggests that κ condition numbers of 30 and higher indicate harmful collinearity, so this was established as a maximum threshold throughout this thesis.

Furthermore, regression models have a number of assumptions which should not be violated. For example, when the outcome variable is linear (i.e. a continuous variable with no discrete levels – e.g. time), it should be guaranteed that this variable is normally distributed (i.e. with most data points being concentrated at the centre of the scale, forming a bell shape). As with most statistical tests, regression analysis also assumes that observations (i.e. values corresponding to individual measurements of a particular variable) are independent from each other. Throughout this thesis it was checked that these assumptions were not violated. The independence assumption, in particular, was handled with the statistical method described below.

3.4.5.2. *Mixed-Effects Models*

In the context of empirical studies on translation or PE, observations are normally interdependent, since various data points usually correspond to a single participant or item (e.g. sentence or text) in the study. It is noteworthy in this respect that repeated measures of this kind usually fail to be properly handled in related research.

In the present study, the interdependence of observations was addressed with the use of mixed-effects regression models (Baayen, Davidson and Bates 2008; see also

Balling 2008), which allow variables in a study to be treated as random effects. Effects classed as random are those that cannot occur more than once in the data and which do not have a fixed range of values or categories – normally participants and items in the context of research in psychology or translation. In the present study, post-editors and ST-MT sentence pairs were random effects in RQ1-2, RQ4 and RQ6, and post-editors, ST-MT sentence pairs, post-edited sentences and quality judges were random effects in RQ5. Treating these variables as *random* renders results more generalizable, as this compensates for any effects driven solely by the items or participants sampled for the study.

Mixed-effects regression models also allow accounting for effects that affect participants in different ways, e.g. some participants may be faster and some may be slower when editing high-quality MT sentences. Effects of this kind can be modelled in mixed-effects regression with the use of random slopes, which act as additional random variables added to the model. This increases confidence that any results obtained are significant over and above variations of the kind described above. Where appropriate, it was also checked in this thesis if random slopes needed to be added to the regression models used.

3.5. Summary

In summary, research questions pertaining to different cognitive aspects of PE are addressed in this study through a mixed-method approach involving two PE tasks: one geared towards the collection of eye movements and subjective ratings on cognitive effort, ‘the eye-tracking task’, and one geared towards the collection of TAPs, ‘the think-aloud task’. Data collected in the eye-tracking task is analysed in its own right whilst also being used as a framework for the analysis of TAPs, with think-aloud results being exploited as a way of expanding on and explaining findings based on eye tracking and subjective ratings. Materials used in these two tasks consist of French news articles with different levels of translating difficulty, corresponding to machine-translated texts whose sentences vary in quality as per values of the automatic evaluation metric Meteor. A total of 28 participants were retained in the study as post-editors: 19 in the eye-tracking task and 9 in the think-aloud task. An additional three participants judged the quality of the raw MT output and the post-edited sentences. The results obtained are presented in the next chapter.

Chapter 4. Results

The results obtained in the present study are reported in this chapter, which is divided into three main parts: results from the eye-tracking task, in section 4.1, results from the think-aloud task, in section 4.2, and an account of the relationship between data obtained in these two tasks, in section 4.3. RQ1 (textual features vs. cognitive effort) is addressed in section 4.1.2, with a more in-depth ramification of the analysis being carried out to address RQ2 (participants' characteristics vs. cognitive effort), in section 4.1.2.1. RQ3 (linguistic aspects of the task attended to) is addressed in section 4.2.2, including details of how participants segmented the task and of the different linguistic aspects of PE they attended to. RQ4 (linguistic aspects of the task vs. cognitive effort) is addressed in section 4.2.4 after a brief explanation of how cognitive effort was estimated in the context of the think-aloud task, in section 4.2.3. RQ5 (effort vs. post-edited quality) is dealt with in section 4.1.3 from a predominantly quantitative perspective, with further qualitative examples being provided in section 4.2.6, based on TAPs. Finally, RQ6 (eye movements and subjective ratings vs. TAPs) is addressed in section 4.3.

4.1. Eye-Tracking Task

4.1.1. *Overall Task Description*

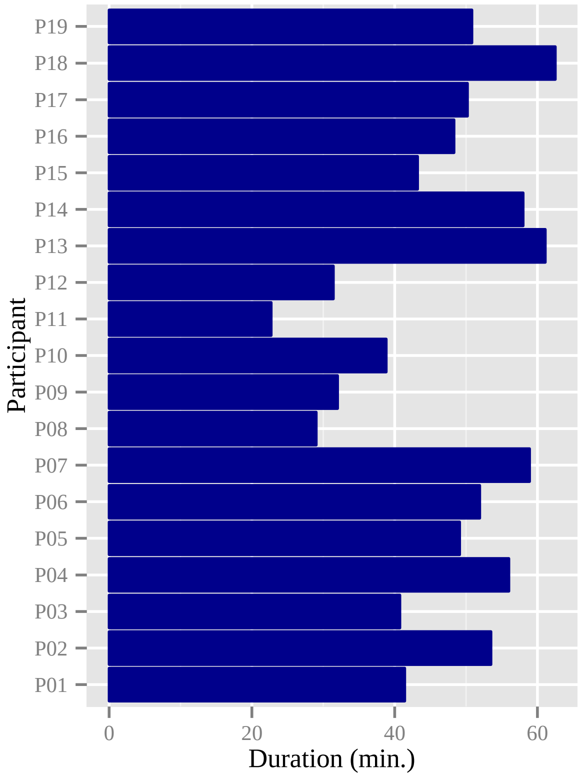


Figure 11 - Per-participant total task time for the eye-tracking task.

An initial observation of the total time spent on the eye-tracking task indicates that there was considerable variation between participants in terms of task time, similarly to previous research in PE (e.g. Guerberof 2014). Total task time including both texts, but excluding additional tasks such as French and working-memory-capacity tests, is presented for each participant in Figure 11. The slowest participant was P18, who took 62.5 minutes to post-edit both texts. The fastest participant was P11, who post-edited both texts in 22.7 minutes, spending just over 10 minutes per text. The average total task time for all participants was 46.3 minutes, with a standard deviation of 11.6.

4.1.2. *Predicting Cognitive Effort*

To address RQ1 and RQ2, the effect of textual features and participants' individual characteristics on cognitive effort was tested based on effort measures previously described in section 3.4.3, namely fixation count, average fixation duration and subjective cognitive effort.

When examining the connection between Meteor's automatic MT evaluation scale and the three measures of cognitive effort considered, it was observed that higher Meteor values (expected to reflect higher MT quality) are associated with less cognitive effort, as would be expected. These relationships are plotted in Figure 12 and Figure 13. Each data point in Figure 12 is a ST-MT sentence pair. As fixation count will inevitably vary with sentence length alone, this measure is presented in the graph per source character. Both measures in Figure 12 were log-transformed, as they did not follow a normal distribution (see Figure 14, below). They were also z-standardised – i.e. being expressed in the same scale, by subtracting the mean and dividing by one standard deviation –, which facilitates a comparison of the effect of Meteor between these two variables. As can be seen in Figure 12 and Figure 13, cognitive effort only decreases with increasing Meteor values up to a certain threshold (approximately 0.6); above that, effort flattens and even increases slightly with rising Meteor scores. In Figure 12, a considerable degree of variation is noted for raw MT sentences with high Meteor scores, which were associated with data points covering nearly the entire range of average fixation duration and fixation count values observed in the study.

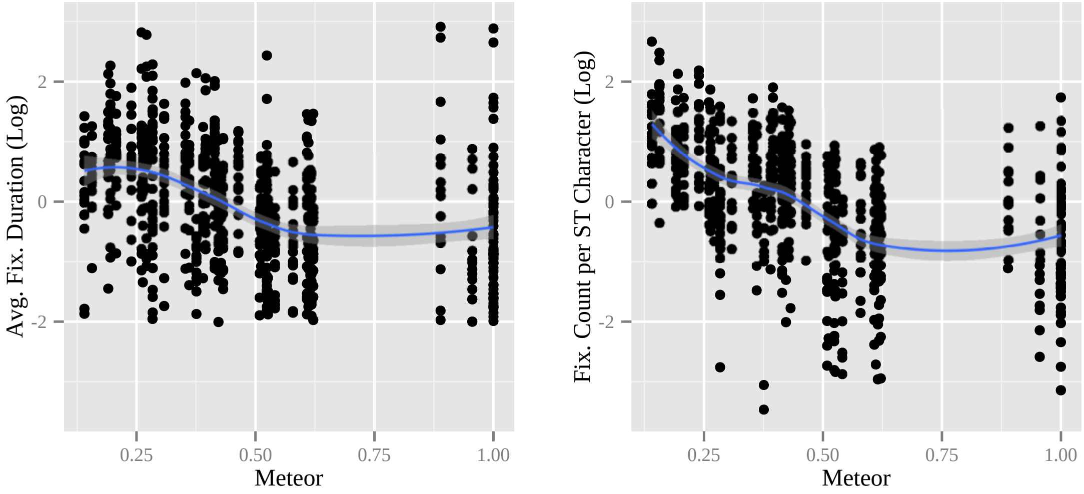


Figure 12 - Meteor (x-axis) and standardised Log Avg. Fix. Duration (y-axis) (left) and standardised Log Fixation Count (y-axis) (right) with loess[44] line and 95% confidence intervals (shades). Fixation count was normalised by ST characters.

[44] 'Loess' stands for local polynomial regression fitting, a regression method (used here for plotting) that makes no assumptions about the data, being able to show any curved relationships.

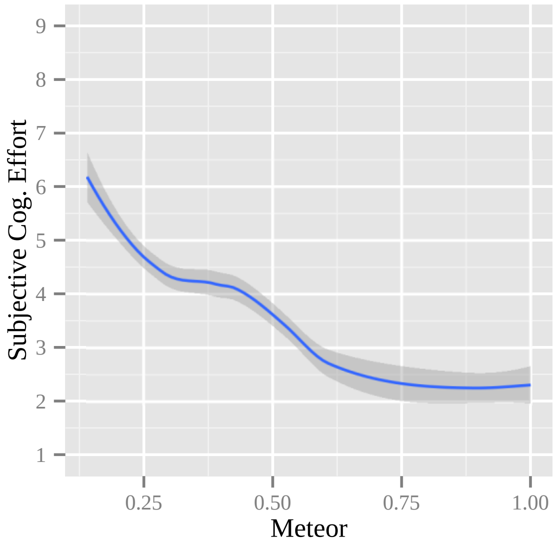


Figure 13 - Meteor (x-axis) and subjective cognitive effort (y-axis), with loess line and 95% confidence intervals.

To further analyse the relationship between these variables whilst taking into account other textual features and participants' individual characteristics, two linear mixed-effects regression models[45] with average fixation duration and fixation count as outcome variables, respectively, were used. Raw values of the latter measures were positively skewed, with most of the data being concentrated on the low side of the scale. Since one of the assumptions of linear regression is that outcome variables are normally distributed (see section 3.4.5.1), these measures were log-transformed, which alleviates this issue. Plots showing the distribution of these variables prior to and post transformation are presented in Figure 14.[46] By their very nature, discrete ranks of subjective cognitive effort are not normally distributed, so they are exploited here in an ordinal model,[47] i.e. a different type of regression model, suitable for ordinal outcome variables.

[45] Fit with the `lmer` function of the `lme4` R package (Bates, Maechler and Bolker 2012).

[46] Using logs of average fixation duration was not found to be necessary for the smaller dataset used by Vieira (2014). As for fixation count, logs of this variable result in a slight negative skew, but this was not deemed problematic as a large amount of the data is concentrated at the centre of the scale after transformation.

[47] Fit with the `clmm` function of the `ordinal` R package (Christensen 2010). Eye-tracking data quality criteria were not taken into account for this model, as perceived cognitive effort is not based on eye tracking.

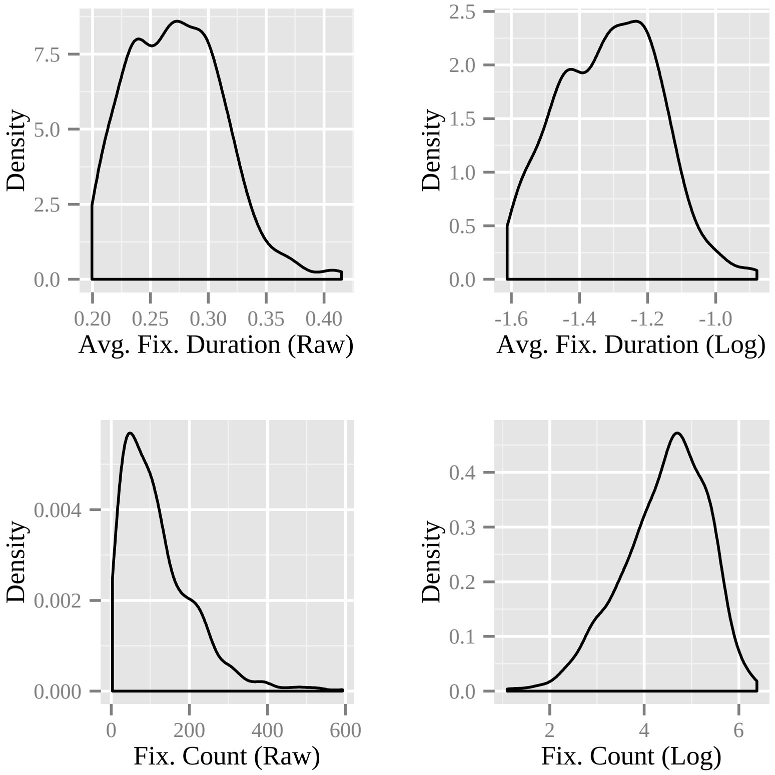


Figure 14 - Distribution of average fixation duration (top) and fixation count (bottom), prior to (left) and post (right) log-transformation.

Spearman's correlation coefficient between scores of subjective cognitive effort and average fixation duration was ρ(706) = 0.40 (p < 0.001), between subjective cognitive effort and fixation count was ρ(706) = 0.67 (p < 0.001), and between average fixation duration and fixation count was ρ(706) = 0.39 (p < 0.001). While moderate correlations between these measures are expected, it is worth noting that two of these correlation coefficients are relatively weak (0.40, 0.39), denoting that these variables could potentially be reflecting slightly different facets of cognitive effort (see sections 4.3 and 6.5 for a detailed discussion on this possibility). With regard to these relationships, it is also important to note that when outcome variables in a study are correlated, multivariate statistical analysis (i.e. when multiple outcome variables are tested concomitantly, in the same model) is a more suitable alternative to running separate models as this avoids redundancy and reduces the risk of false positives. This, however, is only justified when strong correlations are observed (Snijders and Bosker 1999). As analyses of this kind are considerably more complex and as two of the correlation coefficients observed here are rather weak, it seemed desirable to run a separate model for each outcome variable.

Potential predictors included in all three models consist of per-participant French and working-memory-capacity scores, level of professional experience (in years), and per-sentence features listed in section 3.4.1 (see also Appendix G), namely part-of-speech features and seven PCs (principal components) loaded with psycholinguistic and other textual characteristics (see section 3.4.1.3). The number of characters in each source sentence was included as a predictor in all models, accounting for any effects stemming from the length of the sentences alone. Collinearity between per-sentence predictors was measured by inspecting the κ condition number corresponding to these variables (Baayen 2008, 182) (see section 3.4.5.1), which was 9.4, not representing a cause for concern.

Participants and sentences were treated as random effects in all models. Random slopes of Meteor by participant were found to be significant and therefore justified in both linear models.[48] This means that participants reacted differently to Meteor: some may work faster when editing sentences with high scores, for example, while others may spend more time pondering on such sentences, leading to more and/or longer fixations – see section 3.4.5.2 for further information on the use of random slopes.

For simplicity, only one two-way interaction justified by previous research was allowed in the models at this stage (see section 3.4.5.1 for an explanation on interaction effects). Previous research demonstrates that short sentences tend to be penalised by certain automatic evaluation metrics (e.g. O’Brien 2011) and favoured by others (e.g. Green et al. 2014), so it was tested here if the performance of Meteor as a predictor of cognitive effort varied depending on the length of source sentences.

Squared terms of both Meteor and sentence length (i.e. squared versions of these variables) were also included in the models, which is a way of checking for a potential non-linear (i.e. curved) effect of these variables on cognitive effort. These tests seemed justified by plots presented in Figure 12 (p. 99) and Figure 13 (p. 100), where Meteor seems to have a curved effect on cognitive effort, as well as by previous research, where a non-linear effect (represented by the squared term) of sentence length on PE time has been found (Tatsumi 2010).

All predictors were z-standardized prior to entering the models, bringing all variables to the same numerical scale. Initial models with all variables were reduced by (manual) stepwise backwards elimination (Balling and Baayen 2008, 1170) until only

[48] At the time of writing random slopes cannot yet be implemented in the `ordinal` package, hence only included in the linear models. Random effects figures are not reported for economy of space, as these are not the main variables of interest in the analysis.

significant effects ($p < 0.05$) remained. Outlying data points with standardised residuals at more than 2.5 standard deviations from zero were then eliminated and the reduced models were refit (ibid.). P-values for predictors in all models were calculated with the `summary`[49] function.

	Log Avg. Fixation Duration (in sec.)		Log Fixation Count		Subjective Cog. Effort	
Observations	693		691		774	
	β	t	β	t	β	z
Meteor	-0.091	6.52***	-1.188	4.81***	-4.307	5.09***
characters (ST)	-0.01	1.06†	0.373	5.95***	0.479	2.22*
prep. phrases	-	-	0.124	2.43*	0.568	3.04**
PC3	-	-	-0.141	2.73**	-0.498	2.77**
Meteor:characters (ST) (interaction)	-0.042	3.43**	-	-	-	-
Meteor^2 (squared term)	-	-	0.785	3.21**	3.13	3.62***
verb phrases	-	-	-	-	0.392	2.29*

Summary of outcome variables: mean avg. fixation duration 0.271±0.038 SD; mean fixation count 118±91 SD; median subjective cognitive effort 3, mode 3.

Non-significant predictors removed from the models: working memory capacity, French, years of experience, score reflecting attitude towards MT, nouns, verbs, adjectives, adverbs, prepositions, noun phrases, adjectival phrases (all based on the ST); principal components: PC1, PC2, PC4, PC5, PC6, PC7.

KEY:
β = regression coefficients (or ‘estimates’), showing size and directionality of the effect – size can be compared within the same model, but not between models.
t/z = t-value/ Wald’s z-score (higher values stand for lower standard errors, which reflect the statistical accuracy of the estimate).
† Non-significant effect kept in the model as this is part of a significant interaction.
Significance levels: *** $p < 0.001$ ** $p < 0.01$ * $p < 0.05$.

Table 8 - Significant effects in three mixed-effects models predicting average fixation duration (log), fixation count (log) and subjective cognitive effort.

Significant fixed effects that remained in the models are presented in Table 8. Non-significant effects are also listed in Table 8, though no figures are provided as these variables were removed from the models during the fitting stages. As can be seen, Meteor is statistically significant in all three models, with a negative association with cognitive effort (β = -0.091, β = -1.188, β = -4.307). Only Meteor and the interaction between Meteor and sentence length (β = -0.042) remained in the average fixation duration model. This interaction suggests that Meteor is a more accurate predictor of average fixation duration for longer sentences.

[49] Based on the `lmerTest` R package (Kuznetsova, Brockhoff and Christensen 2013) for the linear models.

The squared term of Meteor (Meteor^2) was found to be significant in the fixation count and subjective cognitive effort models. This term denotes a convex curve, with a flatter, upwards tail of the line for sentences with high Meteor scores (see Figure 12, p. 99). It seems that sentences with scores 0.8-1.0 may be perceived as posing as much cognitive effort as sentences with relatively lower scores (e.g. 0.5-0.6), but otherwise Meteor's scale seems generally proportional to effort levels. Indeed, it should be noted that by excluding low-quality MT output from the sample[50] a loss of significance is observed for the squared term of Meteor in the fixation count model. A considerable loss of power is also observed for this effect in the subjective cognitive effort model, though here it remains significant. Checking for a potential outlier effect of these sentences seemed desirable as, in a professional setting, the use of higher-quality MT would normally be aimed for.

PC3, a principal component loaded with variables described in section 3.4.1, was significant in the fixation count and subjective cognitive effort models (β = -0.141, β = -0.498, respectively). Though it is not straightforward to precisely estimate how each individual variable in the principal component affects cognitive effort, by examining PC3's variable loadings, it was noted that source type-token ratio is the feature that contributes most to it, with a positive loading of 0.64 (i.e. of 1.0). Given the negative sign of PC3 in the models, this result suggests that higher type-token ratio is associated with lower levels of cognitive effort. Type-token ratio basically reflects word repetition, with higher values indicating fewer repeated words. While repetition throughout an entire text could arguably be regarded as a facilitating effect, in the present study type-token ratio was measured for each sentence, which suggests that repeated words close to each other hinder the fluency of the text. An example is provided below (Ex.10), where the repetition of words in the French source, such as the expression *à partir de*, is linked to a lack of fluency in the translation.

(Ex.10) **ST**: *Je recommande donc le test à partir de 50 ans, ou à partir de 40 ans si on a un parent direct qui a déjà eu un cancer de la prostate.* ['I therefore recommend the test from the age of 50, or from the age of 40 if one has a close relative who previously had prostate cancer.']

MT: I recommend the test therefore to leave from 50 years, or to leave from 40 years if one has a direct parent who has dj had a cancer of the prostate.

[50] Namely sentences rated with the lowest level by all three quality judges on the scale of either fluency or adequacy.

Here, most participants opted for replacing ‘to leave from’ with more idiomatic expressions such as ‘from the age of’, often also deleting the word ‘years’.

Though the percentage of nouns in the ST is reported by Green, Heer and Manning (2013) as a general linguistic feature for predicting PE time (based on English-French, English-Arabic and English-German as language pairs), results obtained here suggest that this effect does not generalise to French as source language whilst having measures of cognitive effort as outcome variables, as this feature was not found to be significant (see Table 8). This result is consistent with the analysis of textual features carried out by Specia (2011), where nouns had a high correlation with PE effort for English as source language, but not for French. It is not surprising that part-of-speech features may reflect different things in PE depending on the source language, or its family. This could be due to the stylistic preferences of each language; say, with verbs being more common and therefore potentially less effortful to evaluate in English than in romance languages, for example. This may be the reason why, based on the present study, cognitive effort in PE was not found to be associated with ST nouns, but rather with ST prepositional phrases, in both the fixation count and subjective cognitive effort models, and with ST verb phrases, in the subjective cognitive effort model.

Though prepositional phrases are not normally found to have a significant effect on PE time in previous research based on English (see e.g. Aziz, Koponen and Specia 2014; Green, Heer and Manning 2013), this feature is frequently mentioned in the MT literature as a negative translatability indicator (Underwood and Jongejan 2001) (see also section 2.4.1), i.e. a ST feature that is known to pose problems for MT. A source sentence with a high incidence of prepositional phrases is provided below in Ex.11.

(Ex.11) **ST**: *En conséquence, 180 projets de lois restreignant l’exercice du droit de vote dans 41 États furent introduits durant la seule année de 2011.* [‘As a result, 180 bills restricting the exercise of the right to vote in 41 States were introduced in the year 2011 alone’]

MT: As a result, 180 draft laws restricting the exercise of the right to vote in 41 states were introduced during the single year of 2011.

When post-editing this sentence, participants often deleted ‘the exercise of’, opting for just ‘restricting the right to vote’. The time expression at the end of the sentence was also frequently modified, with most participants not deeming ‘during the single year of’ to be idiomatic. One of the reasons for this lack of naturalness in the MT output seems to be the high frequency of the preposition *de* in the French sentence, which is apparently one of the factors accounting for the presence of unnatural English constructions such as ‘the single year of’ and ‘exercise of’ in the translation.

In regard to verb phrases, this effect was not found to be significant by Vieira (2014) as different types of verb phrases were tested separately therein. Here, where all verb phrases were combined into a single variable (see section 3.4.1.1), this feature presents a small effect, in the subjective cognitive effort model. Verbs have been reported as good predictors of PE time by Aziz, Koponen and Specia (2014), in a sub-sentence analysis based on English as source language. In view of results obtained here, it seems that verb-based features are also good predictors for French, though this should ideally be further tested as the effect observed is relatively small.

As for the other variables in the models, an expected positive effect is observed between the length of source sentences and number of fixations landing on the text (β = 0.373), which controls for the impact that sentence length alone has on fixation count. Interestingly, this is also observed for subjective cognitive effort (β = 0.479), which, differently from fixation count, is not expected to vary with sentence length alone. This denotes that participants tend to perceive longer sentences as more challenging, an effect mentioned by Koponen (2012) and also reported to the present author through informal correspondence with professional post-editors.

4.1.2.1. *Participants’ Characteristics and Cognitive Effort*

As can be noted from Table 8, participant variables (working memory capacity, French vocabulary results, score reflecting attitude towards MT and years of professional experience) did not remain in the models as significant effects. Since participant variation is frequently observed in previous research on PE behaviour, it seemed plausible that significant textual effects presented in Table 8 could be moderated by participants’ traits. This was examined with additional tests involving two-way interactions between Meteor and participants’ characteristics. Meteor was chosen for this purpose as it was the variable with the strongest effect in all models.

In the context of these tests, it also seemed interesting to check for any potential effects of the amount of ST use made by participants as this would be expected to be connected to their level of proficiency in French and also to the levels of effort they expended in the task. Eye tracking was exploited to indicate how much attention participants allocated to the ST, which was done by calculating a ratio of the number of fixations that landed on the ST window (see Figure 5, p. 68) over all fixations – henceforth ‘ST fix ratio’. For simplicity, these tests were not carried out in the average fixation duration or fixation count models, but rather only in the model predicting subjective cognitive effort. This avoided the potentially confusing situation of having two eye-tracking measures based on fixations act as both predictor and response variable in the same model. Choosing the subjective cognitive effort model to carry out these additional tests also seemed logical, as participants' individual characteristics would arguably be expected to have a close relationship with the way they perceive the task.

In addition to the interactions between Meteor and subjects' traits, two-way interactions between subjects' traits and ST fix ratio were examined. Here it is important to mention that, interestingly, even participants with low scores in French vocabulary looked at the ST. P03, for instance, a participant who scored 13/100 on the French vocabulary test, had 15% of all fixations landing on the ST window.

As ST fix ratio (i.e. an eye-tracking measure) was being tested as a potential predictor in the subjective cognitive effort model this time, eye-tracking data quality criteria were applied before re-running this model with the interactions, with bad-quality data points being excluded accordingly (see section 3.4.2.3).

Significant results amongst these additional tests are presented in Table 9, which shows that three interactions involving Meteor, level of French, professional experience and ST fix ratio were found to be significant. It should be noted, however, that by excluding low-quality MT sentences from the sample, interactions in the model lose significance. This is nevertheless expected as interactions are more difficult to capture when there is less variability in the sample.

	Interactions in subjective cog. effort model	
Observations	708	
	β	z
French	-0.057	0.28†
ST fix. ratio	0.181	1.79†
experience	-0.227	0.91†
Meteor	-4.23	5.08***
ST fix ratio:French	-0.242	2.55*
French:Meteor	0.294	3.55***
Meteor:experience	0.185	2.43*

All other two-way interactions involving Meteor, experience, working memory capacity, ST fix ratio and score reflecting attitude towards MT were not found to be significant.

KEY:
† Non-significant effects kept in the model as these are part of significant interactions.

Table 9 - Significant interactions involving participants' characteristics in the subjective cognitive effort model.

The interactions involving Meteor suggest that those with more knowledge of French and a larger amount of professional experience reported to expend higher levels of effort on sentences receiving higher Meteor scores (β = 0.294 and β = 0.185, respectively). These effects are arguably expected, as these participants could be deemed more able to identify problems in the translation as they are capable of performing more thorough cross-checks between source and target, being more rigorous with high-scoring sentences.

The interaction between Meteor and knowledge of French is shown in Figure 15. For illustrative purposes, participants were divided into two discrete levels of French knowledge, split at the median of French scores for the sample (89). In view of the odd number of subjects taking part in the study, the participant standing at the median was excluded from the graph shown in Figure 15. The difference between low and high scorers in the French test seems quite small. Nevertheless, it is possible to notice that in the case of low scorers lower-effort levels seem to be associated with higher Meteor scores, which can be observed for effort levels 1, 3, 4, 6 and 8. The opposite is observed for high scorers, who seem to present an association between high-effort levels and higher Meteor scores – the case of effort levels 2, 5, 7 and 9. A wide confidence interval bar can be noticed for level 9 in the case of high scorers. This was due to a single sentence with a Meteor score of 0.89 rated with level 9 on the subjective cognitive effort scale, by P08. Excluding this potential outlier from the analysis did not alter the results.

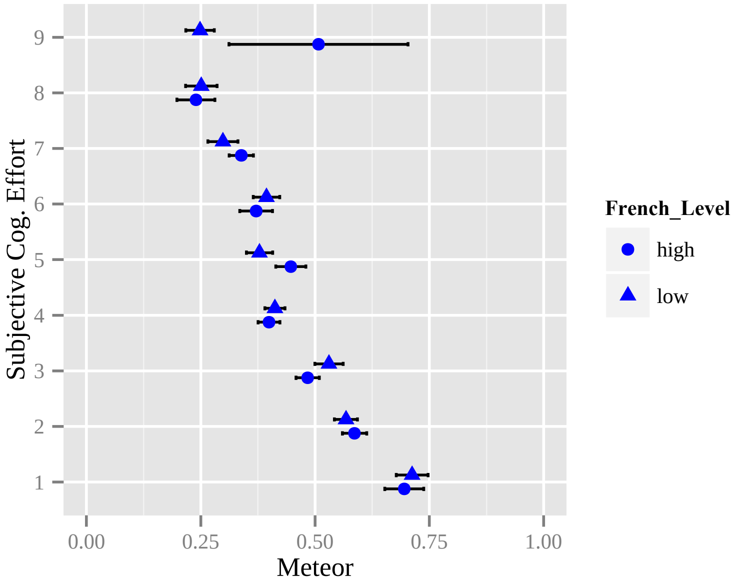


Figure 15 - Average Meteor scores (x-axis) for each level of subjective cognitive effort (y-axis) for low (triangles) and high (points) scorers in the French test.

An interaction between ST fix ratio and French knowledge can also be observed in Table 9 (β = - 0.242). This interaction suggests that consulting the ST is associated with higher subjective cognitive effort for those with less knowledge of French. For illustrative purposes, ST fix ratio plots for subsets pertaining to the 9 top- and bottom-scoring participants in the French test are presented in Figure 16. Similarly to Figure 15, data from the participant standing at the median of French vocabulary scores was excluded from the plots in view of the odd number of participants in the study. As can be seen, increasing ST fix ratio is linked to increasing effort for those in the low-scoring group, with no particular pattern in this respect being noticed for high scorers. Based on these results, it can be tentatively posited that ST fix ratio is associated with subjective cognitive effort mostly in the case of those whose level of source-language proficiency is not very high, as the need to consult the ST might pose more effort to these participants in view of their little knowledge of the source language. In the case of high scorers, subjective cognitive effort does not seem to be related to a higher rate of ST consultation, which, in this case, may reflect mere overall checks not regarded as effortful.

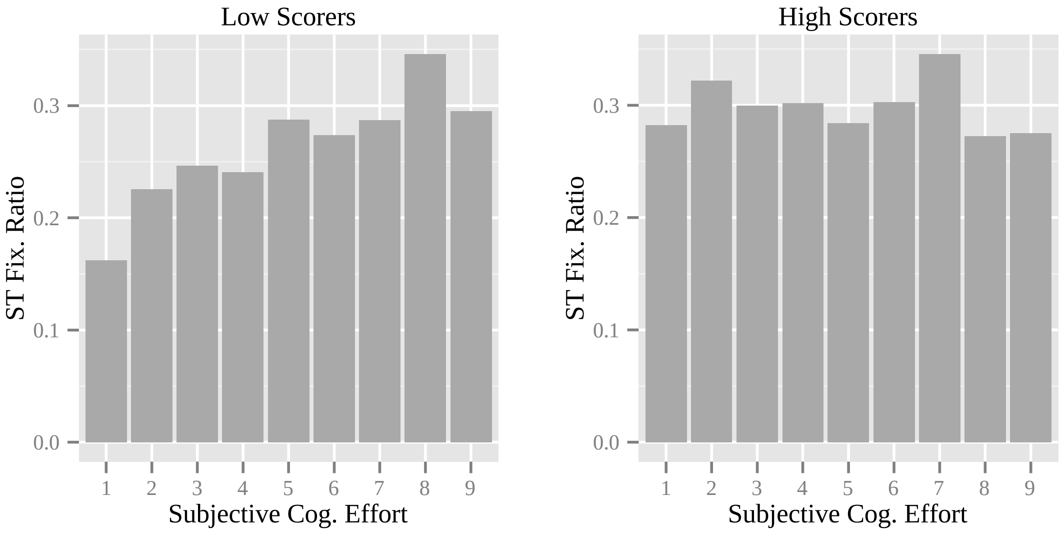


Figure 16 - Relationship between average ST fix ratio (y-axis) and subjective cognitive effort (x-axis) for low French vocabulary scorers (left) and high French vocabulary scorers (right).

Indeed, of sentence pairs that required high rates of ST consultation, highest average ST fix ratio in the case of low scorers is associated with a sentence of low quality, with a Meteor score of 0.16. This indicates that these participants seem to be looking at the ST as a way of solving MT problems, though potentially also having difficulty in interpreting the ST. For high scorers, on the other hand, the sentence with highest ST fix ratio has a Meteor score of 0.61, a relatively high score that suggests that ST consultation in the case of participants with high proficiency in French is not associated with bad-quality MT.

When interpreting these results whilst also taking into account the interaction between Meteor and French knowledge, it seems that higher source-language knowledge allows for a more critical evaluation of the MT output, with effortful situations being centred mainly on the translation itself, as opposed to the ST. The fact that high scorers in French do not present a high ST fix ratio when editing demanding translations is arguably expected as, in the case of these participants, understanding the source should not represent a challenge.

(Ex.12) **ST**: *Aujourd'hui, beaucoup d'hommes chez qui on détecte un cancer ne seront pas traités, car leur cancer n'est pas agressif et ne menace pas leur vie.* ['Today, many men in whom cancer is detected will not be treated, since their cancer is not aggressive and is not life threatening.']

MT: Today, many men in whom cancer is detected will not be treated because their cancer is not aggressive and does threaten their lives. (Meteor: 0.54)

P05, for instance, a participant with a high level of French, rated the MT output provided above in Ex.12 with level 3 out of 9 on the scale of subjective cognitive effort, with 54% of fixations landing on the ST and the MT output being accepted in its original form. The high percentage of ST fixations in this case suggests P05 carried out a number of cross checks between the ST and the MT output before deciding that no modifications were required, a process that still led to a low effort score of 3.

(Ex.13) **ST**: *Au cours des deux dernières années, elles parrainaient des projets de loi dans 34 États pour forcer les électeurs à présenter une carte d'identité avec photo.* ['Over the past two years, they sponsored bills in 34 States to force voters to show a photo ID card.']

MT: Over the past two years, sponsors of the bill in 34 states to force voters to present an identity card with a photograph. (Meteor: 0.43)

For a lower-quality MT sentence, provided above in Ex.13, P05 had a 36% rate of ST use, rating this sentence with level 7 of subjective cognitive effort. P03, a participant with lower proficiency in French, also rated the sentence in Ex.13 with level 7 of subjective cognitive effort. In P03's case, however, 54% of fixations landed on the ST, suggesting that a lot of P03's editing activity involved wrestling with the interpretation of the source. This contrast between P05 and P03 (i.e. participants with high and low proficiency in French, respectively) illustrates results obtained in the regression model, with a higher rate of ST use only reflecting cognitive effort in the case of those with low proficiency in the source language.

Post-edited quality was also inspected for low- and high-French-proficiency groups. Averages comprising all quality judges indicate a fluency score of 3.39/4 and an adequacy score of 3.52/4 for the low French group. The high French group achieved 3.47/4 on fluency and 3.69/4 on adequacy. Results regarding fluency are indicative of very little difference between the two groups (an additional 0.08 for the high proficiency group), suggesting that those with low source-language proficiency can achieve levels of post-edited fluency similar to those achieved by participants with high source-language proficiency. As for adequacy, while results are slightly further apart, relatively little difference is also observed (an additional 0.16 for the high proficiency group). Higher adequacy for the high proficiency group is expected as these participants are better equipped to correct meaning mismatches between source and target. It is interesting to note, however, that levels of both fluency and adequacy are nevertheless broadly similar between the two groups. Per-sentence non-parametric

Mann-Whitney-Wilcoxon tests suggest that any differences in this respect are not significant.[51]

Regarding the results presented in Table 9, it is worth noting that interactions involving working memory capacity did not reach significance and were removed from the model. Tests carried out by Vieira (2014) showed that participants with high working memory capacity were more productive post-editors, but this effect was not found to be significant, which is consistent with the analysis carried out here, based on a larger sample. In view of this, it can be concluded that, despite positive results for this construct observed in other areas (see section 2.2.2.1), working memory capacity seems to be a minor factor in PE. Given the strong theoretical justifications for this effect, however, it seems further investigation in this respect would be of interest, which is discussed in more detail in section 5.1.2.

4.1.3. *Post-Edited Quality*

4.1.3.1. *Fluency and Adequacy Scores*

Despite the diverse nature of the subject sample taking part in the eye-tracking task, relatively little variation was observed between participants in terms of post-edited quality. Average scores for fluency range between 3.22/4 (P11) and 3.56/4 (P05) (sample mean = 3.46/4, SD = 0.4), while average scores for adequacy range between 3.50/4 (P11 and P19) and 3.80/4 (P16) (sample mean = 3.63/4, SD = 0.5). Since all participants were native speakers of English with some experience in translation, it would have been surprising to observe a high number of scores corresponding, for example, to ‘incomprehensible’ or ‘none’ on the fluency and adequacy scales. This narrow range is therefore arguably expected given the scales utilised here.

As judges blindly assessed both the raw MT output and the post-edited versions, it is then possible to examine rates of translation quality improvement across the sample taking different MT quality levels into account. This was done by pooling scores from all judges and checking to see, out of all occasions where a raw-MT sentence was assigned to a specific level (1-4), how many times respective post-edited versions were deemed to be of worse, same or better quality.

These results are presented in Table 10, where it can be seen that, overall, post-edited versions were only deemed to be worse in fluency than the raw MT output on 3% of scoring occasions and in adequacy on 5% (see ‘All’ row) (samples of raw

[51] Adequacy: W = 712.5, p = 0.24; Fluency: W = 708.5, p = 0.22.

fluency and adequacy scores are provided in Appendix G). When looking at figures reflecting overall improvement, on 52% of scoring occasions post-edited versions were deemed better than the MT output in fluency and on 45% of occasions they were deemed better in adequacy. High figures for the 'same' column in Table 10, i.e. where quality was simply maintained at the raw-MT-output level, are largely due to the fact that whenever MT quality was deemed the highest possible (4), the best case scenario was for the post-edited sentence to remain at the same level.

MT Fluency	PE Worse	PE Same	PE Better
Level 1	-	9 (2%)	403 (98%)
Level 2	0 (0%)	25 (5%)	522 (95%)
Level 3	10 (1%)	475 (63%)	272 (36%)
Level 4	61 (10%)	545 (19%)	-
All	71 (3%)	1054 (45%)	1197 (52%)
MT Adequacy	**PE Worse**	**PE Same**	**PE Better**
Level 1	-	18 (5%)	338 (95%)
Level 2	1 (0%)	45 (10%)	387 (89%)
Level 3	9 (2%)	193 (36%)	327 (62%)
Level 4	101 (10%)	903 (90%)	-
All	111 (5%)	1159 (50%)	1052 (45%)

Table 10 -Translation improvement counts showing number of times post-edited sentences were deemed by any of the judges to be of worse, same or better quality in comparison with the raw MT output. Figures between brackets are row percentages.

It is interesting to note in Table 10 that out of raw MT sentences deemed to be at level 3 of fluency, on the majority of scoring occasions (63%) post-edited versions simply remained at the same level, without improving the MT output. This seems to denote a difficulty in improving machine translations standing at an intermediate level of fluency, which could be due to a potential difficulty in correctly identifying or addressing problems in these sentences, expected to be subtler. This is not observed for adequacy, where at all raw-MT-output levels the majority of post-edited versions were deemed to be better than the MT output, excluding those corresponding to MT adequacy of level 4, which could only be deemed the same or worse. This difference between adequacy and fluency in the rate of improvement observed for MT sentences at level 3 could be due to the arguably straightforward nature of adequacy errors, whose solutions would be expected to be clear from the ST on most occasions. This is consistent with the fact that higher inter-rater reliability was obtained for adequacy (see section 3.3.6.2).

Even though quality judges were not specifically asked to indicate errors in the translations to justify their scoring, they had the opportunity of leaving any observations in a comments section available from the evaluation interface (see Figure 9, p. 78). J01, in particular, made frequent use of this section, so this judge's comments are taken into account here to exemplify the nature of frequent fluency and adequacy errors encountered. A supra-sentential error typology analysis was not deemed consistent with the way the eye-tracking task was carried out – i.e. on a sentence-by-sentence basis, with no backtracking – so an analysis of this kind was not implemented. In the context of this task, checking for aspects such as document-level consistency and paragraph development would have arguably been an unfair approach in view of the restrictive operating conditions participants were under. Nevertheless, errors at the sentence level are reflected in the fluency and adequacy scores provided, which is clear from examples of J01's scores and comments, presented in Table 11.

ST	*On va leur suggérer de faire de la surveillance active et si la maladie progresse, on va leur offrir un traitement.* ['Active monitoring will be suggested and, if the disease progresses, they will be offered treatment.']
MT	It will suggest active monitoring and if the disease progresses, it will provide a treatment.
PE	It is suggested to them to actively monitor the disease and if the disease progresses, treatment will be provided.
	MT – Fluency: 2, Adequacy: 2. PE – Fluency: 3, Adequacy: 4. Comment on post-edited version: 'awkward passive voice use'.
ST	*Les dirigeants républicains justifièrent leur politique par la nécessité de lutter contre la fraude électorale.* ['Republican leaders justified their policy by the need to combat electoral fraud.']
MT	The Republican leaders justified their policies by the need to combat electoral fraud.
PE	The Republican leaders justified their policies by stressing the need to combat electoral fraud.
	MT – Fluency: 4, Adequacy: 4. PE – Fluency: 4, Adequacy: 3. Comment on post-edited version: ''stressing' is added to the translation'.

Table 11 - Examples of fluency and adequacy errors encountered based on comments and scores provided by J01 – judges were exposed to pairs including a translation (either post-edited or raw MT) and the French source.

Based on J01's comments, most frequent fluency errors involved problems with punctuation, use of prepositions and articles, verb tense and word order. Most adequacy errors stemmed from verbal constructions judged to modify the meaning of the ST, or concepts/specific words that had been added to or were missing from the post-edited version.

4.1.3.2. *Post-Editing Effort and Post-Edited Quality*

The think-aloud condition has been shown in previous research to incur more controlled cognitive processes in translation, resulting in translated texts of higher quality (Jakobsen 2003). To avoid any potential interference in this respect, RQ5 (i.e. links between PE effort and post-edited quality) is addressed in this thesis based primarily on data collected in the eye-tracking task, allowing a number of different PE effort indicators to be contrasted in view of their relationship with post-edited fluency and adequacy. TAP examples are nonetheless presented in section 4.2.6 in providing a further qualitative account of these relationships.

Effort indicators whose potential links with post-edited quality were examined comprise eye-tracking metrics, PE time, subjective cognitive effort as well as Human-Targeted Translation Edit Rate (HTER), a measure of edit distance that reflects the extent of the modifications implemented in the raw MT output (see section 2.3.2.1).

The eye-tracking measures chosen as potential quality predictors are those exploited as indicators of cognitive effort in section 4.1.2, namely average fixation duration and fixation count. Fixation count was normalised by the number of characters in each source sentence as a way of cancelling any effects caused by sentence length alone. PE time was considered based on the duration of scenes obtained with Tobii Studio. PE time was also normalised by the number of ST characters in each sentence.

In brief, the effect of five different measures of PE effort on post-edited quality is taken into account: fixation count (normalised by ST character), average fixation duration, PE time (normalised by ST character), HTER and 1-9 scores of subjective cognitive effort provided by participants after editing each sentence. These variables can be deemed to capture different facets of the overall concept of PE effort (see section 2.2.1.1). Fixation count, average fixation duration and subjective ratings are, together, regarded here as measures of cognitive effort. PE time represents what Krings (2001) refers to as temporal effort, while HTER provides a rough indication of the levels of technical effort expended by participants.[52]

The relationship between these measures and post-edited quality was analysed with mixed-effects regression models, as described in section 3.4.5.2. Quality judges, post-edited translations, post-editors and ST-MT sentence pairs were treated as random effects in the analysis. Even though scores of fluency and adequacy were found to be slightly correlated ($\rho(2320) = 0.33$, $p < 0.001$), the alternative of running separate models seemed to outweigh a more complex multivariate approach, so two ordinal models[53] were used, having ordered scores of adequacy and fluency as outcome variables.

Measures of PE effort, described above, were tested as potential predictors in both models. Raw-MT-output ratings provided by the judges, as well as post-editors' individual characteristics (namely level of professional experience, working-memory-capacity and French scores, and attitude towards MT) were also included in the models. HTER and the normalised measures of PE time and fixation count were positively skewed, with most of the data being concentrated on the low side of the scale. To minimise a potentially harmful effect of this skew, these measures were square-root-transformed prior to entering the models. For simplicity, subjective cognitive effort and attitude towards MT were treated as numeric (i.e. non-categorical) variables. All numeric variables were z-standardised (i.e. by subtracting the mean and dividing by one SD). Naturally, moderate to high levels of correlation were observed between measures of PE effort, but collinearity was not deemed to be a problem, as the κ condition number (Baayen 2008, 182) for these variables was 10, which can be considered acceptable (see section 3.4.5.1). The models were fit by including new variables and excluding non-significant ones in stepwise fashion. The effect of MT quality and participant variables was inspected first. Once these variables had been accounted for, the impact of different measures of PE effort was tested, also checking for potential non-linear effects of these measures. Significant PE effort variables that remained in the models were also allowed to interact with raw MT quality. These interactions were tested because it seemed likely that the relationship between effort and

[52] Despite the relative simplicity of this score and its widespread use in previous research, this indication is indeed only a rough one, as HTER does not take keyboard use into account – section 2.3.2.1.

[53] Fit with the `clmm` function of the `ordinal` R package (Christensen 2010).

final quality could be moderated by the quality of the existing machine-translation version. Results in both fluency and adequacy models are presented in Table 12.[54]

Fluency			Adequacy		
	β	*z*		β	*z*
MT2	1.5	3.23**	MT2	0.34	1.57†
MT3	2.68	5.02***	MT3	0.78	2.49*
MT4	3.68	6.21***	MT4	1.90	5.38***
HTER	3.81	4.49***	French	0.41	3.12**
HTER2	-1.51	3.12**	Fix	-0.89	6.52***
Fix	-0.72	5.32***	-	-	-
MT2:HTER	-1.05	2.18*	-	-	-
MT3:HTER	-1.83	3.06**	-	-	-
MT4:HTER	-4.02	6.16***	-	-	-

Summary of outcome variables: median fluency: 3/ median adequacy: 4
Number of observations in both models: 2124
Non-significant predictors: level of professional experience, attitude towards MT, PE time, subjective cognitive effort and average fixation duration.

KEY:
MT (2-4): raw MT fluency and adequacy categories/ French: score in French vocabulary test/ Fix: fixation count per character in ST (square root) / HTER: HTER (square root)/ HTER2: squared term of square-root-transformed HTER
† Non-significant term kept in the model as this a two-way comparison (2-1) within an overall significant variable: MT quality.
*** p < 0.001 ** p < 0.01 * p < 0.05

Table 12 - Results of mixed-effects models predicting post-edited fluency (left) and adequacy (right).

As it would be expected, the quality of the raw MT output has a strong significant effect in both models. Since raw MT fluency and adequacy are ordinal scales with four levels (1-4), results for the main effects of these variables (i.e. those that are not part of an interaction) pertain to the difference between each upper level (MT 2-4) and level 1 (the lowest), which functions as a reference. It is observed in Table 12 that there are statistically significant differences when all upper levels of raw MT fluency are compared with level 1 (MT2: β = 1.5; MT 3: β = 2.68; MT 4: β = 3.68). A strong connection between raw-MT and post-edited adequacy is also observed, though here there is no significant difference between MT adequacy levels 2 and 1 (MT2: β = 0.34).

[54] Significance was assessed with the `summary` function and interactions were only kept in the models in case their overall effect was found to be significant based on log-likelihood tests. Raw samples of sentence variables found to be significant in the model are provided in Appendix G.

When comparing MT adequacy levels 3 and 4 with level 1, significant differences exist (MT3: β = 0.78; MT4: β = 1.90).

Another arguably expected effect is the positive relationship between knowledge of French and post-edited adequacy (β = 0.41), which denotes that those with a higher level of proficiency in the source language produced more adequate post-edited texts, with fewer meaning discrepancies between source and target.

In regard to measures reflecting post-editing behaviour, two effects were found to be significant: HTER (i.e. the amount of modifications implemented in the raw MT output), which has an effect on fluency, and the normalised measure of fixation count (Fix), which has a significant effect in both the fluency and adequacy models. The effects observed for these two measures are described below in detail.

In the fluency model, an interaction between HTER and raw MT quality was found to be significant (see three coefficients for MT2-4:HTER). The main effect of HTER in the model (HTER: β = 3.81) holds for MT fluency at level 1, which denotes that at this MT level (lowest quality), more edits (i.e. higher HTER) have a strong positive association with post-edited fluency. The effect of HTER at other levels of raw MT fluency can be observed by subtracting the interaction coefficients (i.e. coefficients for MT2-4:HTER) from the coefficient of the main effect (i.e. 3.81). At MT level 2, HTER also has a positive effect on post-edited fluency. This can be observed by subtracting 1.05 (the coefficient for MT2:HTER) from 3.81, which gives β = 2.76. At MT level 3, HTER has a smaller positive effect on post-edited fluency (β = 1.98, i.e. 3.81 - 1.83). At MT level 4 the effect is reversed, with higher HTER (many changes) being associated with lower post-edited fluency (β = -0.21, i.e. 3.81 - 4.02). HTER also had an overall non-linear effect in the fluency model, as denoted by the squared term of this variable (HTER2: β = -1.51). This suggests that while, overall, modifying the MT output improves post-edited fluency to a certain extent this effect is reversed if many edits are carried out.

Indeed, the interaction effect between HTER and post-edited fluency suggests that the better the raw MT output, the less relevant the impact of changes implemented in the text is. It is noteworthy that this effect is fully reversed at MT level 4, denoting that when MT fluency is the highest possible, changes implemented in the text (when in large quantities) are not indifferent; rather, they are linked to *decreasing* fluency. This is consistent with PE guidelines put forth in the translation industry, which instruct making use of as much of the existing text as possible (TAUS/CNGL 2010). Naturally, this guideline would be expected to guarantee higher productivity, but results obtained here suggest that it may also be associated with higher quality. While the overall instruction of keeping changes to a minimum is already known to the field (see e.g. Guzmán 2007), to the knowledge of the present author this is the first time that empirical evidence supporting a principle of minimal edits is obtained in terms of product quality whilst also drawing a parallel with different levels of raw *MT* quality.

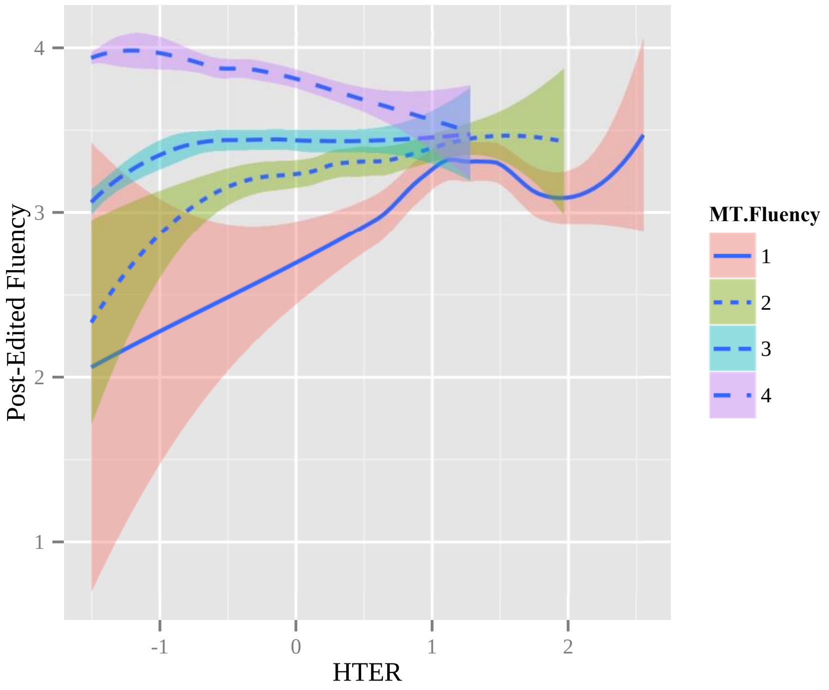


Figure 17 – Loess lines showing effect of HTER (square root, in z-scores) (x-axis) on post-edited fluency (y-axis) at four levels of raw MT fluency – shades are 95% confidence intervals.

The relationship between HTER and post-edited fluency is plotted in Figure 17 considering scores provided by all three judges. Each of the four lines in the graph represents the effect of HTER on post-edited fluency at a given level of raw MT fluency. Shaded areas are 95% confidence intervals, which give an idea of the number of data points supporting the effect: the wider the shades, the fewer the observations driving the line's slope.

As can be seen in Figure 17 by the steep line representing the effect of HTER on post-edited fluency at MT fluency level 1, the raw MT outputs that required the most radical editing were those with the lowest fluency, as expected. Based on modal scores of MT fluency (i.e. the most frequent score provided by the judges for each sentence[55]), it was noted that only one sentence remained at fluency level 1 after PE had been

[55] This was done as a way of disregarding repetitions in the data (i.e. three scores for each translation version) and simplifying any comparisons. When all three judges disagreed, the average was considered instead.

carried out, with few changes being implemented by the participant (P11). The infrequency of cases of this kind, i.e. where poor MT is not improved, is indicated in Figure 17 by the wide confidence interval (shaded area) on the left-hand side of the graph for the MT.Fluency 1 line.

For MT fluency level 2, it was noted that increasing HTER is also linked to increasing post-edited fluency. For MT fluency level 3, a similar effect is observed, though with moderate to large quantities of edits (the x-axis) being indifferent to post-edited fluency. Of raw MT sentences at levels 1-3, most post-edited versions were rated with level 3, though higher HTER was observed for versions that were improved to level 4.

Based on modal scores (see p. 119), it was noticed that most raw MT sentences rated with level 4 remained at the same level after PE, with a low HTER. This means that for most of the time participants managed to maintain the high quality of good machine suggestions. This was only not observed in the case of 14 post-edited sentences, denoting that the negative effect of HTER at MT fluency level 4, mentioned above, is relatively rare. Here, it was noticed that a single sentence fell from MT level 4 to post-edited level 2. A potential outlier effect of this case was inspected by excluding this sentence from the data, but no significant changes could be observed in the results. To illustrate the negative effect of excessive editing observed, the single sentence going from MT level 4 to post-edited level 2 and a sentence going from MT level 4 to post-edited level 3 are provided below in Ex.14 and Ex.15, respectively.

(Ex.14) **ST**: *Passer le test ou non?* [‘Take the test or not?’]

MT: Take the test or not?

PE (P10): Carry out tests or not?

(Ex.15) **ST**: *Dans les études menées aux États-Unis, il y avait beaucoup de contamination entre les groupes témoins, il est donc difficile d’interpréter ces données et d’avoir des recommandations fermes.* [‘In studies conducted in the United States, there was a lot of contamination between control groups, so it is difficult to interpret the data and make firm recommendations.’]

MT: In studies conducted in the United States, there was a lot of contamination between groups, so it is difficult to interpret the data and have firm recommendations.

PE (P13): In studies conducted in the United States, there was a lot of data contamination between groups of those electing to take the test, so it is difficult to interpret this and give firm recommendations.

It is noticeable that the sentence in Ex.14 is quite short. Even though HTER scores are automatically normalised by sentence length, this could be indicating that the score might penalise short sentences, where any absolute amount of changes will inevitably lead to a higher relative proportion of the sentence being modified. However, the effect observed in the model still seems to hold for longer sentences, as illustrated in Ex.15. Interestingly, participants who post-edited sentences in Ex.14 and Ex.15 had prior professional experience and obtained high scores in the French vocabulary test (89 and 97).

An extensive use of the ST together with an overcritical evaluation of the MT output was initially considered as a potential explanation for a detrimental effect of many edits on post-edited fluency. It could be hypothesised that by relying too much on the ST participants could envisage their own preferred translation for the text, which could come into conflict with the existing one provided by the machine, triggering over-criticism. Based on the evidence available from the data, however, this was only partially found to be the case. The rate of ST use associated with decreases in MT fluency was in fact lower than that observed in cases where MT outputs with a high level of fluency simply stayed at the same level after PE. In any way, a conflict between participants' suggestions and the MT output seems indeed to be a factor in cases where decreases of quality are observed. Radical editing of the MT output can be noticed in Ex.15, for instance. In this example, some of the edits carried out by P13 happened in a cascade effect, stemming from the incompatibility of previous edits with the existing text. For example, the participant first opted for 'data contamination' instead of just 'contamination', but the word 'data' already occurred later in the sentence, which made the participant opt for 'interpret this' instead of 'interpret the data' later on. This modification could be deemed not only to account for a certain lack of flow in the sentence but also to add a subtle degree of ambiguity to the text – i.e. does 'this' refer to 'data' or to everything that was previously said? In this respect, it seems that a seamless integration between the raw MT output and modified passages is at the same time a central goal of the task and something that can prove to be challenging even for participants with professional experience in translation and translation revision (the case of P13).

As for the effect of fixation count, a surprising result was observed in the models: fixation count has a negative effect on both post-edited fluency (Fix: β = -0.72) and post-edited adequacy (β = -0.89). These results, which are independent from the quality of the raw MT output, suggest that, while small numbers of fixations on the text are of course required to carry out the task, large numbers are negatively related to the quality of the post-edited product overall – an arguably very surprising finding described below in detail.

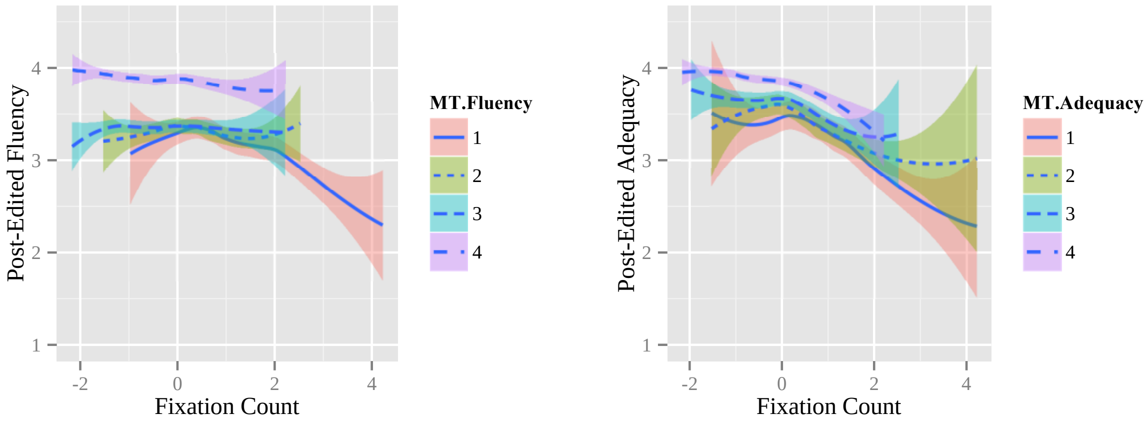


Figure 18 - Non-parametric loess lines showing effect of fixation count (normalised by source characters, in z-scores) (x-axis) on post-edited fluency (y-axis, left) and post-edited adequacy (y-axis, right), at four levels of raw MT quality – shades are 95% confidence intervals.

The effects of fixation count on post-edited fluency and adequacy are illustrated in Figure 18. As can be seen, differently to the effect of HTER, the normalised measure of fixation count tends to present a negative association with the quality of the post-edited text. However, while this is observed at most levels of raw-MT-output adequacy, it is interesting to note that at MT fluency levels 2 and 3 and at MT adequacy level 2, a positive relationship between fixation count and final quality is observed when the number of fixations landing on the text is low. This means that increasing fixation count does often correlate with improvements in post-edited quality (especially in fluency), but as fixation count carries on increasing post-edited quality levels either flatten out or sharply decrease, driving the results observed in Table 12.

While this effect could seem counterintuitive, a number of reasons seem able to explain it. It seems that a large number of fixations could be a sign of deliberation or high levels of difficulty, without a positive effect being noticed on product quality. Based on modal scores (see p. 119), Ex.17, below, shows a sentence that went from raw MT adequacy 4 to post-edited adequacy 3.

(Ex.17) **ST**: *En conséquence, 180 projets de lois restreignant l'exercice du droit de vote dans 41 États furent introduits durant la seule année de 2011.* ['As a result, 180 bills restricting the exercise of the right to vote in 41 States were introduced in 2011 alone.']

MT: As a result, 180 draft laws restricting the exercise of the right to vote in 41 states were introduced during the single year of 2011.

PE (P07): As a result, 180 draft laws restricting the right to vote in 41 states were introduced within a single year - 2011.

In Ex.17, it seems that P07 (who scored 97/100 on the French vocabulary test) attempted to avoid the string 'during the single year of' in the MT output, opting instead for 'within a single year'. The quality judges seemed to think that this altered the meaning conveyed by the original. This indeed seems to be the case: P07's version could imply that draft laws were introduced in any period of 12 consecutive months (i.e. a year), as opposed to during the course of a calendar year (which might not be entirely clear just by the use of the dash before 2011). P07's editing of this sentence involved 180 fixations on the text (both ST and MT/TT), which in comparison with other participants is the largest count for this same sentence (tied with P16).

As the measure of fixation count used in the adequacy and fluency models does not distinguish between fixations landing on the source or target texts, it seemed interesting to inspect if perhaps fixations landing specifically on the source had a different relationship with post-edited adequacy, since high adequacy scores would naturally presuppose an accurate interpretation of the ST. When checking for this, however, the same effects were observed both in terms of absolute counts of ST fixations and of ST fix ratio (the same measure used in section 4.1.2.1, consisting of a ratio of ST-specific fixations over all fixations).

When comparing the effects observed for HTER and fixation count, it can therefore be concluded that, overall, edits have a more positive relationship with post-edited quality than the number of fixations landing on the text, which presented a negative effect. Interestingly, however, HTER only presented a significant effect on post-edited fluency; it was not found to be significant in the adequacy model. This could be due to edits aimed at correcting adequacy problems being minor and therefore not substantial enough to allow a significant relationship between HTER and final adequacy to be observed.

In carrying out additional comparisons to illustrate the contrast between HTER and fixation count, it was observed that out of all machine-translated sentences with a modal fluency score of 2, corresponding post-edited versions that presented no improvement in quality are associated, on average, with an HTER score of 0.22 and with 1.23 fixation per character. When examining successful post-edited versions for these same sentences (i.e. post-edited by other participants) where final quality was improved from a modal fluency score of 2 to a modal score of 4, an HTER of 0.29 and 0.93 fixation per character are observed, on average (i.e. higher HTER and fewer fixations). Fixation count was found to be highly correlated with PE time (ρ(706) = 0.96, p < 0.001), so this seems to indicate that edits carried out under short intervals of time may be associated with higher levels of product quality. As this comparison was done here based on the same MT sentences, edited by different participants, this indeed appears to be an effect of participants' editing behaviour as opposed to just a matter of certain MT sentences being easier to fix.

In Ex.16, below, an MT sentence with modal fluency 2 is presented together with two corresponding post-edited versions that presented discrepant fixation count and HTER values.

(Ex.16) **ST**: *Les nouvelles restrictions affectent de manière disproportionnée les jeunes, les minorités, et les personnes à faible revenue.* ['The new restrictions disproportionately affect young people, minorities and people with low incomes.']

MT: The new restrictions allocate the youths, the minorities, and the people in a disproportionate way to weak income.

PE (P05): The new restrictions disproportionately affect young people, minority groups and people on a low income. (HTER: 0.76; 81 fixations; fluency: 4; 53 seconds)

PE (P07): The new restrictions affect young people, minorities and people of low-income in a disproportionate way to the rest of citizens. (HTER: 0.5; 109 fixations; fluency: 3; 65 seconds)

As can be seen in Ex.16, P07's post-edited version for the sentence presented is closer to the raw MT output than P05's, whose interventions were more substantial, involving the replacement of 'youths' with 'young people', 'weak' with 'low', the deletion of the article before 'minorities' and 'people', as well as alterations stretching across a longer textual span, such as bringing 'disproportionate' to the beginning of the sentence, turning it into an adverb. Indeed, P07's HTER score for this sentence is 0.5 while P05's is 0.76, indicating that P05 modified the raw MT output to a larger extent. P07's post-edited sentence received a modal fluency score of 3, denoting a slight improvement over the raw MT output, while P05's version was rated with a modal fluency score of 4, i.e. the highest possible. Perhaps surprisingly, however, not only did

P05 carry out more changes, but she also did it in less time and whilst fixating less on the text. P05 edited this MT sentence in 53 seconds, fixating her eyes on the text (both ST and MT/TT) 81 times, while P07 edited this same sentence in 63 seconds (i.e. taking over a minute), fixating on the text (both ST and MT/TT) 109 times. These two participants have a relatively similar background (no professional experience and a high level of French), so a direct effect of a quick editing behaviour could be behind P05's good performance on this sentence, a finding discussed further in 5.2.2.2.

As in section 4.1.2, it was also checked if any changes would be observed in the results by excluding low-quality MT from the sample. No significant differences were observed in the models predicting post-edited quality upon excluding raw MT sentences receiving the lowest score of fluency or adequacy, with the trends and effects previously described remaining the same.

4.1.4. *Summary of Results from the Eye-Tracking Task*

Results from the eye-tracking task point to a number of findings regarding predictors of cognitive effort and the relationship between cognitive effort and post-edited quality. As for textual predictors of cognitive effort (RQ1), characteristics based both on the ST and on the MT output were found to be significant. Amongst ST features, type-token ratio and the incidence of prepositional phrases presented the strongest relationship with the levels of cognitive effort expended by participants. The incidence of verb phrases was also found to be a factor, though this effect was smaller, only reaching significance in one of the models used to predict cognitive effort. Amongst MT-output features, the automatic translation evaluation metric Meteor presented an overall strong and highly significant relationship with cognitive effort. This effect, however, was found to be non-linear in two models, denoting that, in the context of the present study, the score functions as a better predictor for values below 0.6. As it would be expected, sentence length had a significant effect on the number of eye fixations landing on the text. Here, results from previous research (Koponen 2012) were replicated, with sentence length also having a small but significant effect on subjective cognitive effort.

Main effects of participants' individual characteristics (RQ2) were not found to be significant in any of the main models used, denoting that when taking all of the previously described factors into account, the effect of textual characteristics on cognitive effort is stronger than that of participants' own characteristics. Nevertheless, when testing for interactions in the subjective cognitive effort model, it was found that for those with a lower level of proficiency in the source language ST consultation was associated with higher subjective cognitive effort. It was also found that those with a high level of source-language proficiency seem to be more critical of translations receiving higher Meteor scores, investing more effort in these sentences. Working memory capacity was not found to be a statistically significant predictor of cognitive effort, but further research on this effect seems of interest to the field.

Ordinal mixed-effects models predicting post-edited quality (RQ5) have shown that larger numbers of eye fixations on the text are not associated with higher post-edited fluency or adequacy. In this respect, interventions in the raw MT output, as reflected by the edit distance score HTER, were found to be associated with higher post-edited fluency, but this is reversed when the fluency of the raw MT output is already high. Further probing of this effect showed that interventions in the text carried out in short intervals of time seem to be the most efficient ones. While these findings seem in line with traditional PE guidelines, to the knowledge of the present author no empirical information in this respect is available in previous research. Other significant effects on post-edited quality include an expected positive effect of the quality of the raw MT output as well as a positive relationship between participants' level of source-language proficiency and the adequacy of the post-edited text.

4.2. Think-Aloud Task

4.2.1. *Overall Task Description*

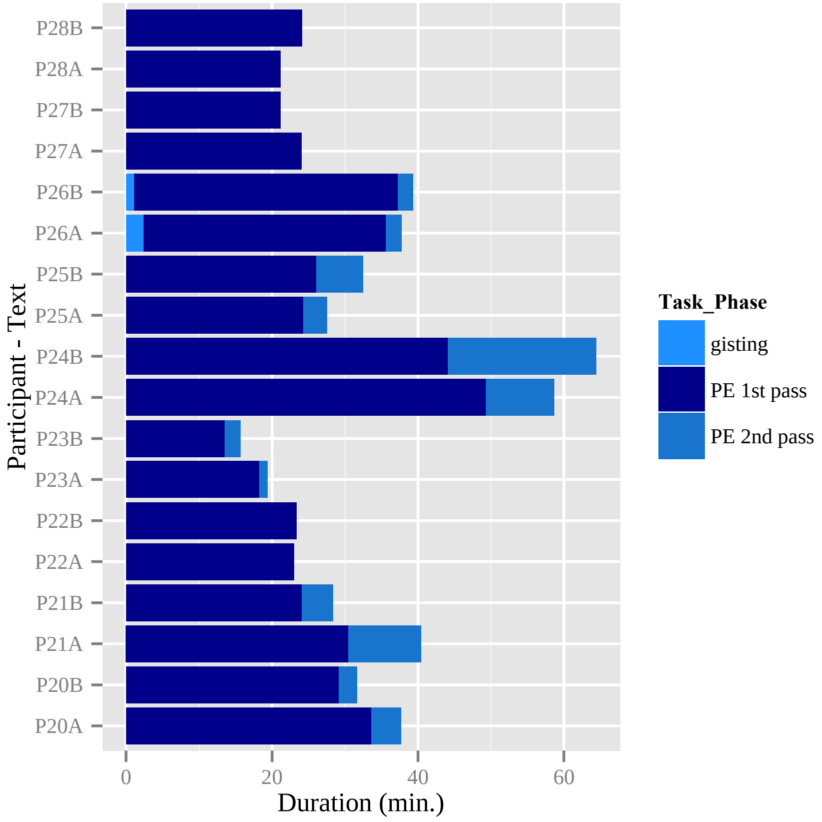


Figure 19 - Total duration of think-aloud task, per participant, per text.

Results obtained in the think-aloud task are reported in this section. Total PE duration for this task is presented in Figure 19 per text for all participants. Three task phases were clear from the TAPs, potentially in line with findings for translation put forth by Carl, Kay and Jensen (2010), who refer to these phases as ‘gisting’, ‘drafting’ and ‘post-editing’, i.e. skimming the text, drafting the translation and revising the draft, respectively. In their formulation, the gisting phase is characterised by reading the text prior to translating, with no use of the keyboard. In the present study, only P26 demonstrated a similar behaviour, with declarations such as ‘I’m going to start by reading the English text’. Most participants then seemed to post-edit the texts in two phases: a first pass where the bulk of the editing was undertaken and a second pass where they checked their work. Though regressions could be observed still within the first pass, a second pass was evident when participants went back to the very beginning of the text and restarted, with statements such as ‘OK, I’m going to have a final read-through’ and ‘in a real-life situation I would go through and check it again’. This was observed for seven out of nine participants.

These phases are further illustrated in Figure 20, which shows keyboarding patterns for four texts post-edited by four archetypical participants. The amount of keystrokes (insertions and deletions, as obtained with Translog-II) carried out by these participants can be observed on the y-axis, with the x-axis representing task minutes. In the case of P26 (top left pane), the initial vertical line indicates the end of the gisting phase, a period where no editing was undertaken and which lasted for approximately two minutes. Vertical lines on the right-hand side of the plots indicate the start of the second pass, when participants went back to the beginning of the texts for another run-through.

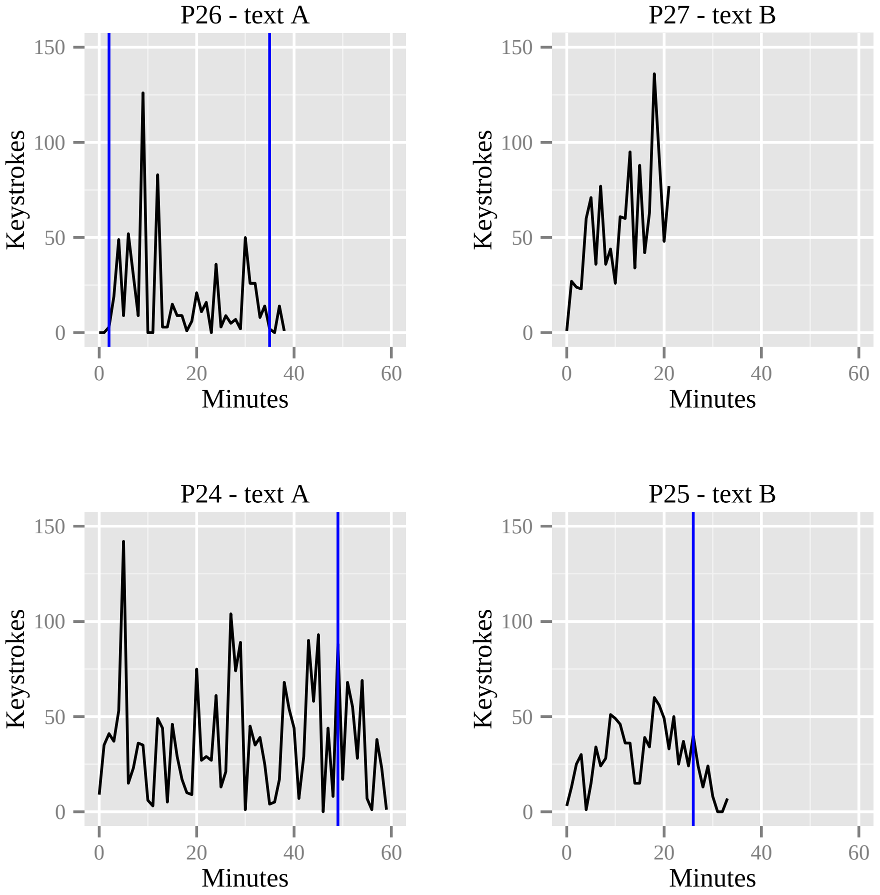


Figure 20 - Amount of keyboarding (y-axis) per minute (x-axis) – vertical lines indicate the beginning and/or end of the first PE pass.

As can be seen in Figure 20, differences between participants in terms of task time and keyboarding behaviour are quite striking. P24 (bottom left pane) took an hour to post-edit text A, performing a substantial amount of changes in between lots of pauses, as indicated by dips in the graph. P24 still engaged in a great deal of editing in the second pass, though going through two entire minutes (approximately minute 56 and 57) consisting mainly of reading. P25 (bottom right pane) had a very dense keyboarding behaviour, with fewer edits being implemented at a time but at a steady rate throughout the task, going through a single minute without implementing any edits whilst still within the first pass. In the second pass, P25 carries out fewer edits overall, with the last minutes of the task consisting almost entirely of reading. P27 (top right pane) was one of the fastest participants overall, not going through either gisting or a second pass. The lack of a final revising stage in the case of P27 is clear from the graph in Figure 20, which ends in an upward trend. P26 (top left pane) carried out relatively fewer edits in the second pass in comparison with other participants who also went through this phase (e.g. P24 and P25). This could be related to the fact that P26 went through a gisting phase at the beginning, which could have enabled this participant to solve a larger number of problems in the text still within the first pass.

In regard to these differences between participants, it seems that the level of professional experience could constitute an influencing factor, though this alone is certainly not able to explain the discrepancies observed. P25 is the most experienced participant in the sample, which could explain this participant's dense editing behaviour with little time needed for thinking. However, P27 was a postgraduate student with no professional experience and also presented a dense editing pattern, having no pauses longer than a minute at any point in the task.

The levels of quality of the resulting post-edited texts are also testament to the complexity of these connections, which is also evident from previous research (de Almeida 2013; Guerberof 2014). P25, for example, had an overall average fluency score of 3.47/4 and average adequacy of 3.74/4. Despite having no professional experience, P27 had an overall average fluency score of 3.56/4 and average adequacy of 3.79/4, both slightly higher than results achieved by P25, an experienced participant.

While Figure 20 sheds light on how participants approached the think-aloud task, an in-depth analysis of keyboarding patterns is beyond the scope of this thesis, with the foci of participants' attention and the levels of cognitive effort they expended being of more direct interest here instead. As for the levels of post-edited quality achieved in the think-aloud task, further details are provided in section 4.2.6. Raw samples of variables exploited throughout sections 4.2 and 4.3 are provided in Appendix G.

4.2.2. *TAP Foci*

To address RQ3 (i.e. linguistic aspects of the task participants attended to), the distribution of TAP categories previously described in section 3.4.4.2 was inspected. Since not all participants went through gisting or a second pass through the text, only TAP units pertaining to the first pass were taken into account for a quantitative analysis.

The distribution of different coding categories, based on both texts and all participants, is presented in Figure 21. When looking at this distribution, it is clear that Reading and Grammar are together the most frequent categories, accounting for approximately 29% of all TAP units each. Lexis closely follows, with 26% of units. These three categories accounted for 84% of all units, denoting that the vast majority of PE mental processes involve non-specific reading/evaluation, or processes focused on lexical or grammatical/syntactical aspects of the activity. The different TAP foci observed in the study are described below.

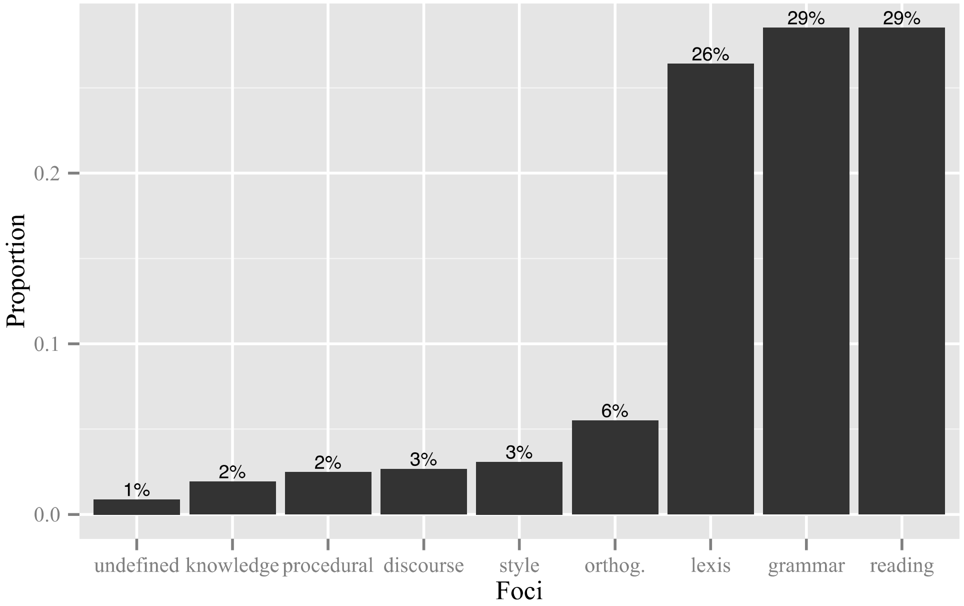


Figure 21 - Distribution of TAP coding categories for entire first-pass phase of the think-aloud task, based on all participants and texts, with rounded percentages.

4.2.2.1. *Non-Specific Reading/Evaluation*

As mentioned in section 3.4.4.2, the Reading category involves mainly the act of reading the text (ST or MT/TT) without a purpose that could be linked to one of the other coding categories. This usually took place when participants were reading a given sentence or text string for the first time or when they had finished editing a given text string and read the resulting TT at the end for a final check. Non-specific evaluative statements were also coded with the Reading category. These comprised, for example, declarations such as ‘urm, that’s difficult’ or ‘yeah, I think that’s OK’. Most Reading units, however, were actual reading events, as the motives behind evaluative statements were usually clear from preceding or succeeding TAP units, allowing them to be coded with a specific category.

4.2.2.2. *Grammar/Syntax*

A wide array of themes could be observed within participants’ verbalisations coded with the Grammar category, including actual grammar errors in the MT output, involving, for example, prepositions and word order, as well as comments pertaining to ST comprehension.[56] An example of a seemingly easy-to-correct error in the MT output was the use of the adjective ‘important’ as a noun in one of the sentences, to which all participants promptly reacted by adding the word ‘thing’ to the translation, as shown below in Ex.18. For most participants, simply reading the words that immediately followed in this instance (‘is to’) seemed to be enough to identify the solution for this problem.

(Ex.18) U916 *But the important... **thing*** [is to have a debate with…] (Grammar)

Other grammar issues seemed more demanding. Ambiguous links between pronouns and referents, for example, often required participants to shift attention to other parts of the text. This was frequently observed in the specific case of an MT sentence where ‘it’ was incorrectly used as a translation for the pronoun *on*, in French, whose closest English equivalents would be ‘one’ (as a pronoun, in the third person singular) or a generic use of ‘we’, as illustrated below in Ex.19.

[56] Exact figures on the number of TAP coding units relating to ST comprehension and to TT production are not provided as this would require extremely fine-grained sub-classifications within each TAP category and also accounting for cases that referred specifically to the transfer between source and target – e.g. ‘I think that’s the original meaning conveyed’ (P23). In view of this, such potential sub-classifications within each TAP category are only qualitatively discussed – see section 2.1 for global figures in this respect observed by Krings (2001).

(Ex.19)	U871	*It will suggest…*	(Reading)
	U872	what?	(Reading)
	U873	I'm gon- *we will suggest to him to do active surveillance* [sight translation]	(Grammar)
	U874	Urm... So...	(Grammar)
	U875	so, *we will suggest* [~~*it will suggest to active*~~]	(Grammar)

4.2.2.3. *Lexis*

While the Lexis category involved issues relating to an adequate transfer from French to English, Lexis units centred specifically on English were also observed, with participants replacing English words not due to outright translation errors, but rather because a more appropriate term was available. This may in part be explained by the high level of PE adopted for the study, with participants being instructed to aim for a text that could be deemed suitable for publication. This could have triggered processes geared towards polishing the text in addition to any processes aimed at correcting more serious errors.

With respect to the above it was interesting to note that certain lexical issues, relating mostly to collocation patterns, were not promptly noticed by all participants. The MT output corresponding to text A, for example, contained the string 'surveillance of the disease' as a translation for the French *surveillance de la maladie*. While *surveillance* works in French in this context, 'to monitor' is a more appropriate verb to be used with 'disease' in this case in English, as shown in TAP units produced by P24, presented below in Ex.20.

(Ex.20)	U2829	*for active surveillance of the disease*	(Reading)
	U2830	I think you probably ***monitor*** [~~surveillance~~] a disease	(Lexis)
	U2831	rather than... *surveilling* it	(Lexis)
	U2832	I'll read this back	(Procedure)

Perhaps surprisingly, however, P24 was one of only seven (out of 28) participants to replace this occurrence of 'surveillance'. When coming across this issue for the first time, P25, for example, declared that she could not think of a better term, even though she was not entirely satisfied with 'surveillance'. The fact that a prompt solution seemed unavailable to most participants in this case suggests that lexical issues of this kind are not necessarily easily dealt with. This could be due to a strong influence of the ST on participants' choices coupled with the borderline acceptability of certain non-idiomatic literal translations, which can make it harder for post-editors to distance themselves from the MT output and suggest an alternative that is more appropriate in the context of the target language.

Still in regard to Lexis, the global distribution of coding categories observed here is testament to the central role of terminology in PE, in line with a pattern that seems to hold across various text genres, with previous research highlighting the importance of lexis in areas as far apart as technical (see Newmark 1988, 152) and poetry (Jones 2011, 129) translation, denoting that this pattern may in fact generalise to most kinds of translating and para-translating activities. The same could not be said of grammar, which seems to play a more prominent role in the specific context of PE. This is most likely because MT produces more grammar errors that require effort to undo (i.e. in PE), differently from self-revision or traditional translation, where serious grammar/syntax errors are arguably less likely.

4.2.2.4. *Style*

Units coded with the Style category constituted a frequent underlying motivation for deletions and substitutions. This was in large part due to word repetition.[57]

(Ex.21)	U261	too many *determines* and *decides*	(Style)
	U1327	just for a different word	(Style)
	U2134	I don't need ~~*results*~~ in there a second time	(Style)
	U2985	yeah, I don't think I need to repeat *test*	(Style)

Ex.21 shows TAP units where a negative effect of word repetition is explicated by participants. This corroborates and explains results obtained in the mixed-effects models presented in section 4.1.2, where high type-token ratios (i.e. less word repetition, in the same sentence) were found to be associated with lower levels of cognitive effort. Machine translations sampled for the present study suggest that state-of-the-art systems are not yet free from awkward word repetition, which would be expected to affect mainly contexts where full PE is carried out, i.e. where the style and flow of the edited text needs to meet human or near-human levels. As shown in section 4.1.3.2, edits involving substitutions of repeated words also carry the risk of unintentionally hindering the flow and/or changing the meaning of the text. It seems

[57] Units that involved the identification of repeated words with a negative evaluation in this respect fell into Style, while units referring specifically to the search for a different lexical alternative were coded with the Lexis category.

that the effort expended by participants in carrying out such edits stems from the need to search for alternatives that are as adequate as the repeated words in front of them. Even though stylistic problems of this kind received little attention in previous PE research, results reported here indicate that in a full-PE scenario these problems are frequently attended to, which is evident both from eye-tracking data and from the TAPs.

4.2.2.5. *Translation Context/World Knowledge*

The least frequent specific TAP category observed in the data was Knowledge, which accounted for just fewer than 2% of all units. The low frequency of this category indicates a low incidence of content editing in the tasks, i.e. little editing where changes in factual information conveyed by the text are necessary. Editing of this kind was nevertheless carried out.

(Ex.22) **ST**: *On peut télécharger ce document (en anglais pour l'instant, une traduction sera offerte sous peu) à cette adresse: http://ca.movember.com/fr/mens-health/prostate-cancer-screening* ['You can download this document (in English for the time being, a [French] translation will be available shortly) at this address: http://ca.movember.com/fr/mens-health/prostate-cancer-screening']

MT: You can download this document (in English for the moment, a translation will be provided shortly) to this address: http://ca.movember.com/fr/mens-health/prostate-cancer-screening

Ex.22 shows a passage, in text A, which refers to a website whose content was in English. Since the PE tasks were carried out from French into English, information regarding the availability of a French translation for the website was deemed irrelevant by most participants, who opted for suppressing the bracketed segment in the MT output altogether, as illustrated by comments presented below in Ex.23.

(Ex.23)	U301	ah we don't need the information about a translation	(Knowledge)
	U947	that's only relevant in the original text	(Knowledge)

This is an example of a situation where the activity may require external knowledge in addition to knowledge of the source or target languages. Other situations where editing of this kind was necessary involved, for example, the substitution of acronyms, which requires knowledge of how specific terms are phrased in different linguistic contexts. Information of this kind was available in the sheet with background information given to participants before the tasks as well as in the text being edited, in earlier or subsequent passages.

4.2.2.6. *Other TAP Categories*

As for other coding categories, Orthography/Typography/Capitalisation TAP units were mostly associated with low-quality MT output containing malformed or misspelt words. Hyphenation was also a frequent issue falling into this category, though this was less prominent than spelling. Though Discourse units were relatively infrequent in the main phase of the task, as shown in Figure 21, they seemed to play a more important role in the second-pass phase – i.e. in the case of participants who went through this stage. The second pass is described in more detail below.

4.2.2.7. *Edits in the Second PE Pass*

Those who engaged in a second pass through the text seemed to have a different approach to the task, not only in terms of their choice of whether or not to go through this phase, but also in terms of editing operations within this pass itself. P20, P23 and P26, for example, seemed to work under a semi-think-aloud condition in this last phase, doing any reading in silence and only verbalising actual changes implemented in the text. P25 did not seem to work under a lower rate of verbalisation in this last phase, but this participant seemed to keep any final modifications to a minimum, as seen in section 4.2.1. P21 and P24, on the other hand, went through considerably longer second passes, at times even reversing modifications they themselves had introduced.

Edits made in the second-pass phase were quite varied, ranging from the correction of typos to more substantial global changes. Previous research suggests that global-level edits in translation self-revision are more prominent towards the final stages of a task (Shih 2006). A similar trend was observed here, with participants paying special attention to text-level consistency and links between clauses and sentences in the second pass. In the case of P25, this tendency also seemed to apply to the interpretation of the ST, with P25 pointing out that the choice of verb tenses was not entirely consistent in the source, requiring that a decision for or against standardisation in this respect be made with regard to the post-edited text, a realisation that had not occurred until this point in the task. P24 makes similar comments, stating e.g. ‘Now that

I’m thinking about it I like the idea of having used *ID* [as opposed to *identity*] throughout’.

In regard to these realisations, Mossop (2007) points out that concomitantly attending to both local- and global-level issues in translation editing and revision can be particularly challenging, so it seems natural that global-level changes are more prominently observed towards the end of the task, when post-editors have acquired a more general view of the text. Results obtained here suggest that this is a behaviour that applies not only to traditional revision, as observed in previous research, but also to PE.

Other common edits carried out in the second-pass phase of the task involve mainly grammar and lexis, following the pattern observed in the main phase. Here it seems again that changes at this final stage could result from participants’ better grasp of the general tone of the text. This was also the case with a few changes relating to style and discourse. For example, on one occasion P20 decided to move an expression to a different place in the sentence to enhance the argument, affirming: ‘having *to some extent* at the end weakens the point’. Similar realisations concerning longer textual spans were also prominent in the case of P25, who declared e.g. ‘I would like to put *in the United States* immediately after *Canada*, to get the contrast’. Word repetition also triggered a number of changes in the second-pass phase, with P24 affirming e.g. ‘I don’t like *results* being repeated’. P25 attended to the same problem, also in the second pass: ‘no repetition of ~~*results*~~ *– false negatives or false positives* [as opposed to *presents incorrect results, with false negative results or false positives*]’.

4.2.3. *Cognitive Effort Clusters*

To explore a potential connection between different TAP categories and the expenditure of cognitive effort in the think-aloud task, data reflecting cognitive effort collected in the eye-tracking task was used to group ST-MT sentence pairs into clusters expected to pose low, medium and high levels of cognitive effort. This was achieved automatically by making use of the K-means clustering algorithm (MacQueen 1967) – available from the Weka Data Mining toolkit (Hall et al. 2009) –, run on per-sentence averages of mean fixation duration, fixation count and subjective cognitive effort. Fixation count was normalised by ST character in view of its high correlation with sentence length, as denoted by mixed-effects results presented in Table 8 (p. 103). Without this normalisation, only longer or shorter sentences could be clustered together, which did not seem desirable. Average fixation duration already is a normalised measure, since it consists of total fixation duration divided by fixation count, so it would not be logical to normalise this measure again. Subjective cognitive effort, originally a discrete ordinal score, was not normalised either, though for the purpose of clustering it is assumed as a numeric variable. Details of the clusters obtained are presented in Table 13, which also includes per-cluster averages of human-assessed raw-MT-output fluency and adequacy.

	Low Cog. Effort (180 ST words)	Medium Cog. Effort (279 ST words)	High Cog. Effort (385 ST words)
avg. fixation duration (in sec.)	0.243	0.262	0.293
subjective cog. effort (1-9)	2.19	3.38	4.94
fixation count (per ST character)	0.47	0.73	1.25
MT fluency (1-4)	3.6	3.24	1.65
MT adequacy (1-4)	3.87	3.5	1.92

Table 13 - Per-cluster averages of cognitive-effort measures and raw-MT-output quality scores – cognitive effort figures comprise all 19 subjects taking part in the eye-tracking task; quality values take all three judges into account.

As can be seen, cognitive effort measures increase from the first ('low') to the last ('high') cluster. Unsurprisingly, the opposite pattern is observed for human-assessed MT fluency and adequacy scores, with MT sentences that pose lower cognitive effort being of higher quality on average.

4.2.4. *Different TAP Foci and Cognitive Effort*

By taking the clusters described above into account in the analysis of TAP units, it is then possible to check to see if differences can be noticed in the distribution of TAP categories pertaining to passages in the text expected to pose different levels of cognitive effort.

Table 14 shows TAP category distributions and total number of units per cluster.[58] It was noticed that total TAP unit counts increase together with the expected level of cognitive effort posed by sentences in each cluster. This indicates that, despite previous criticisms, TAPs may indeed have the potential of reflecting cognitive effort in PE, with the amount of think-aloud data produced by participants mirroring cognitive effort measures traditionally used in cognitive psychology, such as eye movements and subjective ratings, a finding that is further explored in section 4.3.

[58] TAP units were assigned to their corresponding ST-MT sentence pair based on the object of participants' attention in each specific unit, as indicated by their verbalisations.

	Low Cog. Effort units (/ST word)	Medium Cog. Effort units (/ST word)	High Cog. Effort units (/ST word)
Lexis	89 (0.5)	345 (1.24)	883 (2.29)
Grammar	82 (0.45)	317 (1.14)	934 (2.43)
Style	29 (0.16)	10 (0.04)	111 (0.29)
Orthography	22 (0.12)	40 (0.14)	180 (0.47)
Discourse	14 (0.08)	30 (0.11)	73 (0.19)
Knowledge	27 (0.15)	14 (0.05)	24 (0.06)
Reading	252 (1.4)	375 (1.34)	733 (1.90)
Procedural	8 (0.04)	12 (0.04)	67 (0.17)
Undefined	2 (0.01)	10 (0.04)	25 (0.06)
Total	*525 (2.91)*	*1154 (4.14)*	*3030 (7.87)*

Table 14 - Incidence of different TAP foci over three sentence clusters expected to pose low, medium and high levels of cognitive effort – figures between brackets are TAP unit counts normalised by number of source words in each cluster.

Regarding the distribution of TAP foci in each cluster, it can be observed that an overall increasing pattern is stronger in the case of Lexis and Grammar, denoting that not only are these frequent categories overall but also that they seem to have a connection with the degree of cognitive effort expended by participants. Reading presents a less clear pattern in this respect, with the low- and medium-cognitive-effort clusters having a nearly equal incidence of Reading units when results are normalised by cluster size (i.e. by diving TAP units by ST words).[59] This seems due to the fact that simply reading and/or evaluating the text is expected to be necessary under all circumstances, i.e. with the act of reading the text without a specific problem-solving purpose not being particularly linked to cognitive effort expenditure.

As for other TAP categories, it is noteworthy that the low-cognitive-effort cluster has the largest number of Knowledge units. In this respect, it should be pointed out that the majority of these units correspond to the same sentence: the ST-MT sentence pair presented above in Ex.22 (p. 134). This suggests that the incidence of this category in the low-effort cluster is a rather specific effect, with only 12 Knowledge units remaining in this cluster if the sentence pair in Ex.22 was to be removed from the data.

The number of TAP units coded with the Orthography category increases especially sharply from the medium- to the high-cognitive-effort cluster. Upon

[59] The units were normalised per word and not per character as it did not seem logical that TAP units would vary as a function of word length.

inspection of the sentence pairs driving this effect it was noticed that the vast majority of Orthography units in the high-cognitive-effort cluster correspond to low-quality machine translations containing malformed or non-translated words that required corrections in spelling, which indeed would only be expected in contexts involving low-quality MT.

When results are normalised by cluster size, Style, Orthography, Knowledge and Reading present non-linear patterns across the three clusters, suggesting that these TAP categories have a less appreciable relationship with cognitive effort overall. It may be that aspects other than effort have a stronger link with some of these categories, a possibility discussed further in section 5.1.3.4.

To further investigate patterns observed in Table 14, mixed-effects regression models were used[60] to measure how cognitive effort levels correlated with the occurrence probability of different TAP categories. Mixed-effects modelling is in this case a statistically superior technique as compared, for example, to a chi-square cross-tabs test, which would not be able to handle individual differences between participants. The mixed-effects analysis was carried out through binary comparisons, by checking to see where the effect of cognitive effort was stronger: for specific TAP categories or for Reading, regarded here as a relatively neutral category. Reading seemed like a good parameter for comparison since, as previously mentioned, simply reading/evaluating the text could be regarded as a requirement of the task itself, being necessary irrespective of the level of cognitive effort expended. Potential effects of participants' individual traits were also checked for in the models, as it would be expected that the nature and amount of TAP units produced by participants could vary as a function of their individual characteristics alone.

Each TAP unit was a data point in the analysis. Specific predictors in the regression models consisted of cognitive effort (a three-level categorical variable – see Table 13, p. 137) and per-participant control variables presented in Table 4 (p. 75), namely level of professional experience, and scores reflecting French proficiency and attitude towards MT[61] – all expressed as z-scores (i.e. in the same scale), by subtracting the mean and dividing by one standard deviation.

Table 15 shows significant results ($p < 0.05$, based on the `summary` R function) obtained in the models, where positive coefficients indicate a higher occurrence probability for units coded with specific TAP categories as opposed to Reading.

[60] Fit with the `glmer` function of the `lme4` R package (Bates, Maechler and Bolker 2012).

[61] This score was treated here as a numeric variable, for simplicity.

Procedural and Undefined units were not analysed. All pairwise comparisons of cognitive effort were tested[62] and this variable was only kept in the models if its overall effect was significant.

	Lexis		Grammar		Knowledge	
Observations	2677		2693		1428	
	β	z	β	z	β	z
cog. effort 2 - 1	1.31	3.25**	0.92	4.41***	-	-
cog. effort 3 - 1	1.78	4.62***	1.40	7.16***	-	-
cog. effort 3 - 2	0.47	1.45†	0.48	3.1**	-	-
experience	0.31	1.99*	0.20	2.27*	0.67	2.49*
	Style		Discourse		Orthography	
Observations	1510		1477		1602	
	β	z	β	z	β	z
cog. effort 2 - 1	-1.54	-2.71*	-	-	-	-
cog. effort 3 - 1	0.1	0.23†	-	-	-	-
cog. effort 3 - 2	1.64	3.24**	-	-	-	-
attitude	-0.45	-3.16**	-	-	-	-
experience	-	-	0.66	3.13**	0.27	2.44*

Cognitive effort is presented in terms of between-level comparisons. Subjects and items are treated as random effects in all models.

KEY
†Kept in the model as these are two-way comparisons within a single categorical variable: cognitive effort.
* = p < 0.05; ** = p < 0.01; *** = p < 0.001

Table 15 - Results in mixed-effects binomial models comparing specific linguistic TAP foci with Non-Specific Reading/Evaluation.

It is clear from Table 15 that a higher level of professional experience is associated with a higher occurrence probability for Lexis (β = 0.31), Grammar (β = 0.20), Knowledge (β = 0.67), Discourse (β = 0.66) and Orthography (β = 0.27) units when compared to Reading. An opposite trend occurs for participants' attitude towards MT in the case of Style (and Style only), where the negative coefficient (β = - 0.45) suggests that a more positive attitude towards MT is associated with a lower occurrence probability for Style units.

[62] A higher number of significance tests leads to a higher chance of false positives being obtained. Tukey's test was used here to carry out multiple pairwise comparisons between different levels of cognitive effort whilst compensating for this inflated risk.

In regard to the above, it is interesting to note that the frequency of all specific TAP categories is associated with participants' level of experience except for Style, which seems more strongly associated with a negative attitude towards MT. Also, results suggest a non-linear effect between Style units and cognitive effort. Comparatively to Reading, a significant decrease in the occurrence probability for the Style category was observed from cognitive effort level 1 to level 2 (β = -1.54), while a significant increase can be noticed from level 2 to level 3 (β = 1.64). This effect seems to be due to the discrepant number of Style units in the clusters, as previously observed in Table 14 (p. 138), where it was noticed that just 10 units fall into the medium-cognitive-effort cluster, while 110 units fall into the high-cognitive-effort cluster. Upon inspection, no particular connection was noticed between the 10 units falling into the medium-cognitive-effort cluster, though many of the TAP units falling into the high-cognitive-effort cluster concern participants' choices regarding an age expression in one of the sentences, with participants pondering on whether 'aged', 'years', or 'years old' were necessary words when expressing someone's age. Overall, there seems to be a wider phenomenon of Style working differently when compared to other categories. A larger sample would be needed to shed further light on the relationship between Style and cognitive effort, however, as a more substantial number of Style units would ideally be required for the non-linear effect observed to be confirmed.

As for other specific TAP categories, increasing cognitive effort is significantly linked to an increasing occurrence probability for Grammar and Lexis units, as can be observed in Table 15, which shows positive coefficients comparing different cognitive effort levels with respect to their effect on the occurrence probability of the Lexis and Grammar categories. In the case of Grammar, a significant gap exists between all levels of cognitive effort. In the case of Lexis, an overall significant pattern is also observed, but the gap between cognitive effort levels 2 and 3 was not found to be significant (β = 0.47). This gap is also smaller in the case of Grammar (β = 0.48), though here it is significant. Most importantly, it can be observed that as cognitive effort increases from level 1 to level 3, the occurrence probability for units in the Lexis and Grammar categories also significantly increases when compared to Reading (β = 1.78, β = 1.40).

Still in regard to the Lexis and Grammar categories, Table 15 shows that regression coefficients were more significant for Grammar than for Lexis. However, just comparing coefficients in the two models is not straightforward as these models are based on different overall numbers of observations. In view of this, to compare the effect of cognitive effort between Lexis and Grammar, a different mixed-effects regression model was used checking to see if the effect of cognitive effort levels was significantly stronger for one of these two categories. Interestingly, results in this respect indicate that the difference between Lexis and Grammar is not statistically significant.[63] This is arguably a surprising finding as it suggests that the inappropriacy of MT lexis suggestions is as big a challenge as errors in MT grammar. It should be mentioned in this regard, however, that the limitation imposed here on the use of reference materials cannot be excluded as influencing these results, i.e. by potentially making lexical issues harder to deal with. This influence was nevertheless not deemed to be significant as participants were instructed to trust the MT output whenever they were unable to at least infer the meaning of French words. In addition, a substantial number of Lexis units involved deliberations on the suitability of relatively simple English alternatives which would not be expected to require external consultation, e.g. ‘talk’ vs. ‘discussion’.

4.2.5. *TAP Unit Sequences*

TAP units were not found to have a 1:1 relationship with problems participants came across in the text or aspects of the task they attended to, so a further grouping of the data into higher-level sequences seemed desirable. For this purpose, the data was grouped into micro TAP sequences, regarded here as series of adjacent TAP units that reflect how participants segmented the text, with each sequence corresponding to a single text segment put in working memory for processing. A similar approach is adopted by Jones (2011) in the analysis of TAPs in translation, but micro sequences themselves are coded in his study, as opposed to each individual unit.

In the present study, micro sequences often comprised more than one TAP category. This was the case, for example, when a grammatical change was immediately followed by further lexical adaptations in the same MT output segment, or when a change in the text and an accompanying comment made by the participant were deemed to reflect different foci. In sequence (S) 912, below (Ex.24), after reading a segment in the MT output (‘the test of APS’), P26 first tackles a problem regarding factual information conveyed by the raw MT, i.e. an incorrect acronym. This is evident in units 4251 and 4252. Immediately after correcting the acronym, P26 also makes a modification in structure, opting for using ‘PSA’ as a modifier. In the approach adopted

[63] Figures are not presented for economy of space.

here, all these units were deemed to be part of the same sequence, as they all refer to the same text segment initially put in working memory for processing.

(Ex.24)	S912	U4249	I’ll have a look at the next one	(Procedure)
		U4250	urm, OK, *the test of APS*	(Reading)
		U4251	so I guess this is referring still to the same test	(Knowledge)
		U4252	so it should be ***PSA*** [~~*APS*~~]	(Knowledge)
		U4253	urm and I’m going to call it the *PSA* ***test*** [~~*test of*~~]	(Grammar)

On a number of occasions, participants initially read longer segments but then addressed different problems within these segments one at a time, which often required subsequent re-readings so that shorter and more specific segments were available in working memory. In these cases, each separate smaller segment was regarded as a separate micro sequence, with the act of reading the larger initial segment being included in the first sequence of the series. In cases where a participant read an entire sentence for the first time and moved on with no changes being made, this single reading event was regarded as a separate sequence. When, however, this sentence was re-read as part of a regression triggered by a problem that lay ahead in the text, the reading event was included in the sequence associated with the problem that triggered the regression.

Groups of non-adjacent overlapping sequences could also be observed, i.e. sequences taking place at different task moments and corresponding to the same text segment, henceforth ‘macro sequences’. This occurred, for example, when participants returned to a given problem or part of the text at a later moment in the task. For most participants, however, this was only prominent when taking the second-pass phase of the task into account. As not all participants went through this phase, macro sequences are only qualitatively described here. Results regarding micro sequences for the first-pass phase of the task are presented below.

	Low Cog. Effort	Medium Cog. Effort	High Cog. Effort
Avg. sequence length in TAP units	3.89 (2.76 SD)	4.42 (2.98 SD)	4.68 (3.32 SD)
Total sequence count (/ST word)	133 (0.74)	263 (0.94)	651 (1.69)

Table 16 - TAP sequence count and average sequence length for each cognitive effort cluster.

The average length of micro sequences (in TAP coding units) was 4.4 overall. Table 16 shows average TAP sequence length (with standard deviations) and total sequence count (together with counts normalised by ST word) per cognitive effort cluster. Total counts of micro sequences have a clear linear relationship with cognitive effort levels. A similar pattern occurs for average sequence length, but with cluster figures being extremely close to each other, with high standard deviations. This suggests that high levels of cognitive effort are associated mainly with a larger number of operations (i.e. TAP sequence count). Any potential association between cognitive effort and *longer* operations is, at best, a secondary effect. To investigate what could be influencing the length of operations, TAP sequence length was further examined as a function of different TAP foci.

The incidence of different TAP foci within the bottom and top quartiles (25%) (according to length) of micro sequences is presented below in Table 17, where it is clear that Lexis and Orthography are the categories whose incidence differs most between short and long sequences. The effect of Orthography is arguably expected, as it would be unusual for spelling corrections to require long sequences of TAP units. As for Lexis, while an incidence of 18% is noticed in the bottom quartile, the presence of Lexis in the top quartile is nearly twice as large (30%), denoting that long sequences are disproportionally lexis-based. Perhaps interestingly, no particular effect in this respect was noticed for Grammar, which has a slightly lower incidence in the top quartile.

	Bottom Quartile (short sequences)	Top Quartile (long sequences)
Lexis	18%	30%
Grammar	31%	27%
Style	2%	3%
Discourse	6%	2%
Orthography	10%	3%
Knowledge	2%	2%
Reading	26%	29%

Table 17 - Different TAP foci within short and long sequences.

To further examine these relationships whilst controlling for participant and item variation, a *poisson* mixed-effects model was used[64] with sequence length (in TAP units) as response variable, also testing cognitive effort (i.e. a three-level variable) as a potential predictor. Per-participant control variables presented in Table 4 (p. 75) were also included in this model. The frequency of Grammar, Lexis and Orthography units in the sequences was included in the model as predictors, expressed as proportions varying between 0 and 1. Results suggest that, indeed, Lexis has a significant positive relationship with the length of sequences ($\beta = 0.226$, $z = 4.39$, $p < 0.001$). A significant effect was also found for Orthography units, which had a negative impact on sequence length ($\beta = -0.410$, $z = 4.20$, $p < 0.001$). No significant relationship was observed for Grammar. In line with figures presented in Table 16 (p. 143), cognitive effort levels also failed to have a significant relationship with the length of sequences. In this respect, significant results obtained for Lexis suggest that the very nature of the problems dealt with has a more appreciable relationship with the length of operations (in terms of TAP sequence length) than does the level of cognitive effort expended by participants.

(Ex.25)	S286	U1217	*and twenty per cent of the electorate between*	(grammar)
		U1218	rather than ~~*in*~~	(grammar)
		U1219	***between*** *eighteen* [~~*to*~~] ***and*** *twenty-nine* [~~*years*~~]	(grammar)
	S488	U1012	*it is in this spirit that a majority of governments*	(reading)
		U1013	***American*** *governments*	(grammar)
		U1014	rather than *governments* ~~*American*~~	(grammar)
(Ex.26)	S346	U1583	[*the important*] *thing is to have a debate with*	(lexis)
		U1584	I think *have a discussion*	(lexis)
		U1585	rather than *debate*	(lexis)
		U1586	*have a...*	(lexis)
		U1587	or *talk* maybe [~~*discussion*~~]	(lexis)
		U1588	*have a talk*	(lexis)
		U1589	I'll say ***talk***	(lexis)
		U1590	sounds better	(lexis)
		U1591	*have a talk with the doctor to decer* determine if they should pass him or not*	(reading)

[64] *Poisson* models are normally used for count data. The `glmer` function of the `lme4` R package (Bates, Maechler, and Bolker 2012) was used to fit the model, and the `summary` function was used to obtain the p-values. Participants and items were treated as random effects in the model, which also included a dummy variable at observation level to address over-dispersion – i.e. when the error variance is larger than the mean (see Baayen 2008, 297), a situation that requires special treatment.

These results are illustrated above in Ex.25 and Ex.26. Ex.25 shows two three-unit micro TAP sequences involving changes in preposition and word order, respectively. In Ex.26, a nine-unit sequence is shown involving the replacement of ‘debate’ with ‘talk’. These examples seem to illustrate how decisions of a lexical nature are associated with longer processing sequences, while this may not necessarily be the case for decisions involving grammar. This could be due to the non-human (i.e. simple and obvious) nature of certain MT grammar errors (such as the wrong position of the adjective in S488), which are arguably not expected to require long deliberations to be corrected. However, because they occur in large quantities, producing various micro sequences, grammar errors have a strong association with cognitive effort, as denoted by results presented in Table 15 (p. 140). As for the Lexis category, the very nature of lexical decisions are arguably prone to longer sequences of thought processes, with mutually exclusive and overlapping lexical possibilities being considered on a paradigmatic axis of linguistic analysis. Grammatical aspects, on the other hand, are dealt with in a concatenating fashion, on a syntagmatic axis, producing different sequences as new segments are put into working memory for processing. Another possibility in this respect is that lexical decisions may involve considering more possible candidates than grammar decisions, on average, i.e. the fact that lexis might present participants with a wider array of choices.

With respect to macro sequences, i.e. groups of non-adjacent overlapping micro sequences, substantial variation could be observed, with their occurrence proving to be considerably more frequent for certain participants. This was the case for P21 and P24, for example, who even within their first pass seemed to constantly move backwards and forwards in the text, often taking more than a single micro sequence to solve problems they came across. P20, P25 and P26, on the other hand, seemed to adopt a more linear segmentation strategy, often dealing with problems within micro sequences themselves, without revisiting the corresponding part of the text later on. P20, P25 and P26’s higher level of professional experience could be one of the factors influencing this effect. In the case of P26, in particular, this linear approach to the task seemed to have been strategically planned, with declarations such as: ‘my strategy for this one is going to be to do it in sections, because it is a bit of a longer text’ (text A); ‘I’m going to do it section by section again’ (text B).

For P20, P25 and P26, macro sequences can hardly be observed if the second-pass phase of the task is not taken into account. Here, there is arguably room for further research regarding potential connections between second-pass edits and cognitive effort. An in-depth analysis in this respect is not carried out here in view of the small number of participants going through a second pass and in view of differences between those who did, with some subjects seeming to work under a semi-think-aloud condition in the second pass, as mentioned in section 4.2.2.7.

4.2.6. *Post-Edited Quality*

While, as previously mentioned, the quality of the post-edited text is investigated in this thesis mainly based on data obtained in the eye-tracking task, overall averages of post-edited quality based on the think-aloud task denote that, as expected, raw MT is substantially improved in this task too. Average post-edited fluency for this task was 3.47/4 (0.14 SD) and average post-edited adequacy was 3.65/4 (0.20 SD), against an average raw MT fluency of 2.67/4 (1 SD) and an average raw MT adequacy of 2.93/4 (1.05 SD). It should be noted, however, that an in-depth analysis of post-edited quality based on the think-aloud task would require global aspects of the texts to be taken into account, such as document consistency and paragraph development. Such an analysis could not be carried out in the context of this thesis, but think-aloud data is nonetheless exploited here as a way of shedding further light on quantitative results obtained in the eye-tracking task.

Particularly interesting for the questions addressed here is checking to see how participants' post-editing behaviour, as reflected in the TAPs, relates to the quality of the post-edited product. Here, again, cases where raw MT quality decreased could be observed.

(Ex.27) **ST**: *Or, le Centre Brennan considère cette dernière comme un mythe, affirmant que la fraude électorale est plus rare aux États-Unis que le nombre de personnes tuées par la foudre.* ['However, the Brennan Centre considers this a myth, stating that electoral fraud is rarer in the United States than the number of people killed by lightning.']

MT: However, the Brennan Center considers the latter as a myth, stating that electoral fraud is rarer in the United States than the number of people killed by lightning.

TAP units:

U4510	*However, the Brenan Center considers the latter as a myth*	(Reading)
U4511	we don't we wouldn't need the ~~*as*~~ in English	(Grammar)
U4512	*which considers the latter a myth*	(Reading)
U4513	*stating that electoral fraud is rarer in the United States than the number of people killed by lightning*	(Reading)

U4514	I’m not very keen on the phrasing there	(Grammar)
U4515	urm, I don’t think we can say *rarer than the number of people killed by lightning*	(Grammar)
U4516	*Electoral fraud is rarer in the United States than* [~~*the number of people*~~] ***being*** *killed by lightning*	(Grammar)

Ex.27, above, shows a raw MT sentence rated with level 4 on adequacy and 3 on fluency (based on modal scores) with corresponding TAPs produced by P26. The adequacy of the post-edited sentence in this case was rated with a modal score of 3 while its fluency was given a modal score of 4, denoting an increase in fluency but a decrease in adequacy. As can be seen, changes implemented in this sentence comprise the deletion of ‘as’ in the string ‘considers as’ (U4511) and the substitution of ‘the number of people killed by lightning’ with ‘being killed by lightning’ (U4516). P26 seemed to stumble upon a mismatch concerning the use of ‘rarer’ as a modifier of ‘number of people’. In attempting to address the issue, the modification introduced by P26 drops the notion of quantity, implying arguably more directly than in the original that electoral fraud in the United States is rarer than the likelihood of any individual being killed by lightning, as opposed to only individuals in the United States. While this is ambiguous in the ST (i.e. deaths by lightning in the United States or in general?) and while P26’s suggestion would arguably work well as a standalone version, it is interesting to note regarding this MT sentence that less invasive edits performed by other participants seemed to be more successful in improving fluency whilst maintaining a high level of adequacy. P24, for example, who maintained raw MT adequacy at level 4, opted for the use of ‘rate of electoral fraud is lower’, instead of ‘rarer’, an alternative that recycles a larger amount of the raw MT output by retaining the string ‘number of people killed by lightning’. This again goes to show that a minimal editing approach may be a key factor in avoiding accidental quality losses in PE.

Ex.28, below, shows a raw MT sentence rated with a modal score of 2 on fluency and 3 on adequacy, both being improved to level 4 after PE – in this case, by P24. Here it seems that, differently to that observed above in Ex.27, changes implemented in the text were aimed at addressing issues that come across as being more evidently wrong, such as the use of ‘are’ instead of ‘have’ (U3370). It appears that, in cases of this kind, correcting the raw MT output is more straightforward, since participants change the text with a view to allowing comprehensibility as opposed to achieving slight improvements in the way the text sounds. In the process of achieving such slight improvements, the efficacy of different editing alternatives may be less evident, potentially requiring more effort and accounting for a higher risk of decreases in quality.

(Ex.28) **ST**: *Il est important de noter que, contrairement au Québec, les citoyens américains ne disposent pas de carte d'identité universelle comme la carte de l'assurance maladie.* ['It is important to note that, unlike in Quebec, American citizens do not have a universal ID card such as the health insurance card.']

MT: It is important to note that, unlike Quebec, American citizens are not universal ID such as the health insurance card.

TAP units:

U3367	*Il est important de noter que, contrairement au Québec, les citoyens américains ne disposent pas de carte d'identité universelle comme la carte de l'assurance maladie*	(Reading)
U3368	*It is important to note that* [looks at FR]	(Reading)
U3369	*unlike* ***in*** *Quebec*	(Grammar)
U3370	*au Québec*	(Grammar)
U3370	*American citizens*	(Reading)
U3370	that doesn't make any sense	(Reading)
U3370	oh, *not universal ID* [looks at FR]	(Lexis)
U3370	*have*? *Have no,* ***have no*** [~~are not~~] *universal ID*	(Lexis)
U3370	*dispose* sounds a bit fancier than *have*	(Lexis)
U3370	but that is just French sometimes [laughs]	(Lexis)
U3370	[looks at FR and EN] *have no universal ID…* [looks at FR] *such as the health insurance card*	(Reading)

Based on similar examples found in the data, post-edited results from the think-aloud task seem again to denote that traditional PE guidelines that recommend keeping changes to a minimum and using as much of the existing text as possible indeed represent the most efficient way of approaching the activity. However, results obtained in this thesis suggest that gauging what 'minimum' stands for can represent a challenge. This seems to be especially the case when the quality of the MT output is of a medium level or already relatively high, a situation where seamlessly combining the existing text with any edits may prove difficult.

4.2.7. *Summary of Results from the Think-Aloud Task*

Results from the think-aloud task shed light on a number of aspects of PE that would be hard to investigate by making use of automatic logging methods alone. At a deeper level than just considering modifications implemented in the text or typical MT errors, it was observed in this task that the vast majority of linguistic aspects of PE (RQ3) concern either simply reading/evaluating the text or addressing grammatical and lexical aspects of the activity. Though other aspects such as world knowledge or discourse were also attended to by participants, these aspects were comparatively less prominent. The large incidence of Lexis units was regarded as particularly interesting in this respect, as this pattern seems to hold across various translating modes, including traditional translation and PE.

It was also observed based on TAPs that participants segmented the task into three distinct phases: an initial gisting phase, followed by two rounds of PE. Similarly to results observed for translation, not all participants went through gisting or the second PE round.

When drawing a parallel between TAP categories and cognitive effort (RQ4), it was found that both Lexis and Grammar were associated with the amount of effort expended by participants, with higher levels of cognitive effort correlating with a higher occurrence probability for TAP units of a lexical or grammatical nature. In this respect, though the effect observed was slightly stronger for Grammar, the difference between these two categories was not found to be significant, suggesting that grammar and lexis pose very similar levels of cognitive effort in PE.

A significant effect of individual characteristics on TAP category probabilities was also found, with more experienced participants presenting an overall tendency towards producing TAP units with a specific linguistic focus, as opposed to simply reading the text without explicating a motivation or problem-solving objective.

4.3. Think-Aloud Data as a Measure of Cognitive Effort

This section aims at addressing RQ6 (i.e. eye movements and subjective ratings vs. TAPs) by checking to see if significant links can be observed between data obtained in the eye-tracking task and spoken data produced by participants under a think-aloud condition. As previously mentioned, participant samples in both tasks are comparable in terms of French knowledge, previous experience and attitude towards MT (see section 3.3.5).

4.3.1. *Verbalisation Effort and Cognitive Effort Clusters*

A frequent finding from TAP-based studies is that the think-aloud condition entails a slow-down effect (Jakobsen 2003; Krings 2001). It is not frequently mentioned in previous research, however, if and how this effect and other known weaknesses of

TAPs (see section 2.2.3.3) influence distributions within a single dataset, i.e. if the interference posed by the method is appreciable also in relative terms and not only in terms of the total amount of data collected. To examine this possibility, measures gathered in the context of the eye-tracking task were contrasted with think-aloud data with a view to checking to see how TAPs compare with other data sources as potential indicators of cognitive effort. In this respect, a first sign that TAPs may mirror other measures is the fact that a linear relationship was observed between TAP unit counts and cognitive effort clusters (see Table 14, p. 138), i.e. discrete sentence categories representing 'low', 'medium' and 'high' cognitive effort, a classification based on average fixation duration, fixation count and scores of subjective cognitive effort, all based on the eye-tracking task – see section 4.2.3.

The TAP coding units adopted here can be regarded as minimal meaning-bearing chunks within participants' verbalisations. The process of segmenting raw verbalisations into these units is expected to cancel out any effects that may stem simply from the fact that certain participants might use more words than others to convey a similar message. To check if differences in this respect could be observed in the results, the relationship between cognitive effort clusters and TAPs was examined also in terms of the raw number of words verbalised by participants. This is illustrated in Figure 22, with boxplots showing the distribution of absolute number of verbalised words (right) and TAP coding unit count (left) per sentence in each cognitive-effort cluster. As the amount of data produced would be expected to vary with sentence length alone, both these measures were normalised by the word count of each source sentence.

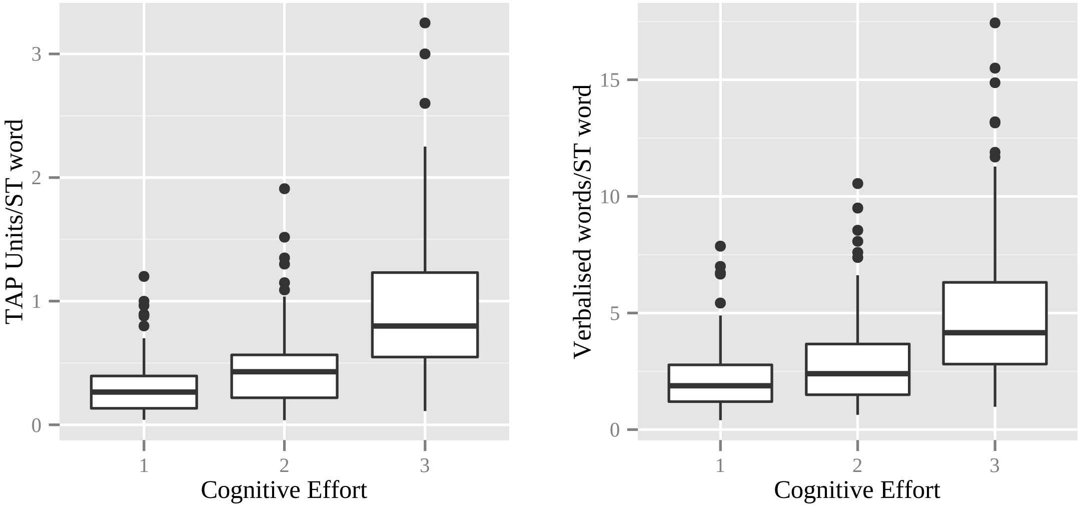


Figure 22 - Boxplots showing the relationship between cognitive effort clusters (x-axis) and rate of verbalised words per source word (y-axis).

Figure 22 shows a clear positive trend both for TAP units and verbalised words, with larger amounts of TAP-based data being observed per sentence in the medium- and high-cognitive-effort clusters, as previously shown in Table 14 (p. 138) in terms of TAP unit count. This suggests that the level of noise potentially introduced by the use of absolute verbalisation measures does not appreciably affect results. Indeed, TAP units and verbalised words were found to be highly correlated ($\rho(367) = 0.93$, $p < 0.001$).

In this respect, it is noteworthy that coding units and verbalised words were found to have different correlations with MT quality by Krings (2001) – while raw verbalisation was linearly connected with quality, a non-linear relationship was observed for coding units, which occurred in larger quantity for medium- as opposed low-quality MT in his study. A potential reason why TAP units and a raw measure of verbalised words produced similar results here lies in the fact that, differently to that done by Krings, most conjunctions were treated as coding unit delimiters in the present study (see section 3.4.4.1), which seemed to allow for a more precise coding of the data based on the coding categories adopted.

It is relevant to note, however, that even in terms of the raw number of words verbalised, a more straightforward metric not affected by segmentation rules, a clear link can be observed between TAPs and other measures of cognitive effort exploited for clustering. To further examine this connection, measures of cognitive effort gathered in the eye-tracking task are explored individually below in section 4.3.2. In view of the high correlation observed between TAP unit count and number of words verbalised, only TAP units are considered hereafter.

4.3.2. *Correlation between TAPs and Cognitive Effort Measures*

To further explore the relationship between TAPs and other measures of cognitive effort, rather than making use of discrete clusters, potential effects of the original variables used for clustering were examined separately. This allows more in-depth information to be revealed with regard to the specific relationship that each measure of cognitive effort considered – namely, fixation count, average fixation duration and subjective cognitive effort (see section 4.2.3) – may have with think-aloud data. This was done in the context of a mixed-effects model, with per-sentence TAP unit counts obtained in the think-aloud task being assumed as the outcome variable and per-sentence averages of cognitive effort measures based on the eye-tracking task being tested as predictors. As more TAP units would be expected to be produced with text length alone, source-sentence length was also included in the model as a predictor. As it is also possible that participants' individual characteristics may influence the amount of TAP units they produced, per-participant variables, namely score in French test, level of professional experience and attitude towards MT, were also included in the model. The model assumed the *poisson* distribution (suitable for counts) whilst also correcting for over-dispersion (see footnote 64). All numeric predictors were z-standardised (i.e. by subtracting the mean and dividing by one standard deviation). The κ condition number for per-sentence variables was 8, so collinearity was not deemed a problem in the model (see section 3.4.5.1 for an explanation on the function and interpretation of the κ condition number).

A comprehensive model with all predictors described above was reduced by backwards stepwise elimination (Balling and Baayen 2008, 1170) – see section 4.1.2. Significant variables that remained in the model are presented in Table 18, which shows that subjective cognitive effort and average fixation duration have a significant positive relationship with TAP unit count (β = 0.415 and β = 0.198, respectively). Per-participant random slopes of both these variables were found to be significant.[65] This means that participants varied significantly on these variables, which has been accounted for in the analysis. The fact that sentence length was found to be significant (β = 0.310) denotes that, as expected, participants produced more TAP units when post-editing longer sentences.

	TAP Unit Count	
Observations	369	
	β	z
sentence length (ST)	0.310	6.90***
subjective cognitive effort	0.415	5.40***
average fixation duration	0.198	2.56*

Non-significant predictors: French vocabulary, attitude towards MT, level of professional experience, fixation count.
* p < 0.05, ** p < 0.01, *** p < 0.001

Table 18 - Significant variables in *poisson* mixed-effects model checking the effect of individual measures of cognitive effort on TAP unit count.

[65] Based on log-likelihood tests.

Of the two cognitive effort measures with significant effects, subjective cognitive effort was the variable presenting the strongest correlation with TAP units (β = 0.415). This correlation is arguably expected since both TAPs and subjective ratings rely on participants' perceptions of the task. In fact, it seems that TAPs and subjective ratings are able to capture slightly different facets of cognitive effort when compared to more objective measures, with subjective measures being potentially more holistic, as mentioned in section 2.2.3.2. It is interesting to note, however, that even a measure based on a relatively more objective method such as eye tracking had a significant correlation with TAPs (i.e. in terms of TAP coding unit count).

These results suggest that, for the purpose of measuring cognitive effort, any interferences between the think-aloud condition and cognitive processes are only significantly large in absolute terms. In investigations where relative comparisons within a single dataset are of interest, these interferences did not prevent a clear link between TAPs and other cognitive-effort measures from being observed. It is in any way worth remembering that a number of procedural differences exist between the two tasks conducted here, namely the fact that participants were only allowed to backtrack in the think-aloud task and that they were only asked to provide subjective ratings on cognitive effort in the eye-tracking task. This, however, is not deemed to invalidate the results as differences of this kind would be expected to incur larger discrepancies between the tasks and not induce any correlations between them.

Indeed, links observed between the eye-tracking and think-aloud tasks suggest that results based on gaze data and subjective ratings on cognitive effort seem to have suffered little interference from the more restrictive operating conditions adopted for the eye-tracking task. When examining the relationship between Meteor and number of TAP units produced in the think-aloud task, for example, striking similarities can be noticed in comparison with the relationship observed between Meteor and eye-tracking measures. This is illustrated in Figure 23, where a flattening effect of the line for high Meteor values is also observed, in line with results obtained in the context of the eye-tracking task, presented in Figure 12 (p. 99).

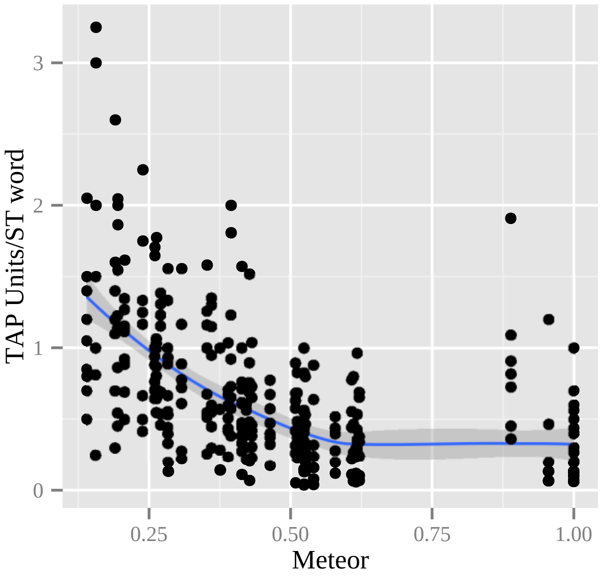


Figure 23 – Loess line showing correlation between Meteor (x-axis) and TAP unit count/ST word (y-axis) – shades are 95% confidence intervals.

With respect to these findings, it should be pointed out that an in-depth investigation of how the think-aloud condition changes the nature of cognitive processes could not be carried out here, as this would ideally require a qualitative analysis of the data obtained in the eye-tracking task similar to the one carried out here based on TAPs. In this respect, previous research in translation has shown that participants tend to break the text into smaller segments when thinking aloud (Jakobsen 2003). However, results obtained here suggest that effects of this kind would be expected to remain relatively constant, without a significant interference being noticed in studies carried out from a relative perspective, based on a single dataset.

4.3.3. *Summary on Links between the Eye-Tracking and Think-Aloud Tasks*

When contrasting data collected in the eye-tracking and think-aloud tasks, striking similarities were observed. Sentence clusters reflecting different levels of cognitive effort based on measures obtained in the eye-tracking task presented a positive relationship with think-aloud data both in terms of number of think-aloud coding units and the raw number of words verbalised by participants. When checking which specific measures of cognitive effort used for clustering seemed to have a more direct relationship with think-aloud unit counts, subjective cognitive effort ratings were the variable with the strongest effect, with a significant effect also being observed for average fixation duration. Despite previous criticisms, these results present empirical evidence suggesting that think-aloud protocols can be regarded as a measure of cognitive effort in PE.

Chapter 5. Discussion

This chapter presents a discussion on results obtained in the two PE tasks conducted in the study. Connections between findings deriving from the eye-tracking and think-aloud tasks are exploited here whenever possible in examining the underpinnings of the phenomena observed and how these relate to the wider context of research in PE and related areas. Implications of the findings obtained for professional practice and for further research are presented in the context of each research question addressed in the investigation, including aspects relating to the prediction of cognitive effort in PE (section 5.1), the relationship between PE effort and post-edited quality (section 5.2), and a comparison between TAPs and other types of data that can be used as measures of cognitive effort (section 5.3).

5.1. Indices of Cognitive Effort in Machine-Translation Post-Editing

5.1.1. *Cognitive Demands of the Text (RQ1)*

The first research question addressed in this thesis (RQ1) concerned the identification of textual characteristics (pertaining either to the ST or to the MT output) that could act as predictors of cognitive effort in PE. Cognitive effort, a construct estimated here through eye-movement data and subjective ratings, is regarded in this thesis as the amount of mental resources individuals expend in response to the cognitive demands of a task. Potential effort predictors tested in the study include the automatic MT evaluation metric Meteor and linguistic characteristics of the ST, such as word frequency and the incidence of different part-of-speech categories. Previous research in this respect lacks a comprehensive approach, with studies related to this thesis (e.g. Aziz, Koponen and Specia 2014; Lacruz, Denkowski and Lavie 2014; O'Brien 2011) evaluating characteristics of *either* the ST or the MT output as predictors of effort. Most of these studies estimate PE effort based on absolute PE time or amount of editing, measures found to be misleading indicators of the overall amount of effort required by the activity (see section 1.1). In response to these limitations, this thesis investigates effects of ST *and* MT-output features on *cognitive* effort, which is regarded in previous research as the 'decisive variable' underlying an overall notion of PE effort (Krings 2001, 179).

In the context of RQ1, five variables were found to be significantly associated with the levels of cognitive effort expended by participants: Meteor, the length of source sentences, a principal component (see section 3.4.1.3) loaded mainly with source type-token ratio – i.e. the amount of word repetition in source sentences –, the incidence of prepositional phrases in the ST and the incidence of verb phrases in the ST. These effects are discussed below in detail.

5.1.1.1. *Meteor*

As an automatic MT evaluation metric, Meteor has presented good correlations with human assessments in previous research (see e.g. Callison-Burch et al. 2012; Machácek and Bojar 2013), but to the knowledge of the present author this score has not been previously tested in the context of PE as a predictor of cognitive effort. Scores used in a similar context in other studies, most frequently TER (Snover et al. 2006) and GTM (Turian, Shen and Melamed 2003), rely only on edit distance, failing to indicate similarity (between the raw MT output and a human reference translation) at a semantic level, a known issue of traditional metrics addressed by Meteor.

O'Brien (2011) tests the performance of GTM as a predictor of eye-tracking measures in PE, but she arbitrarily assumes three bands of the GTM metric as potentially predicting 'low', 'medium' and 'high' levels of cognitive effort. In regard to these three categories, while it is straightforward to assume that a score of 0.8 is numerically higher than a score of 0.5, for example, it cannot be guaranteed that this numerical difference will be proportional to the amount of effort associated with these scores, which renders making a priori assumptions on 'low', 'medium' and 'high' levels an imperfect approach for the purpose of effort prediction.

In the present study, a wide range of Meteor scores was tested, with no discrete categories being established a priori and with other linguistic features (such as type-token ratio and the incidence of different part-of-speech categories in the ST) being controlled for in the analysis. Results in this respect showed that scores higher than 0.8 were at times associated with as much cognitive effort as lower scores, e.g.

0.6.[66] In practical terms, this means that, as an index of cognitive effort, Meteor is more reliable when outside the range of extremely high values, a tendency indicated by the squared term of this variable (i.e. the squared version of the variable itself), found to be significant in two of the mixed-effects models presented in section 4.1.2 – see Figure 12 (p. 99). This non-linear (i.e. curved) effect could be explained by the fact that some participants were found to be more critical than others when editing sentences receiving high Meteor scores, as seen later in section 4.1.2.1. Individual differences of this kind could have produced the curved effect observed, with the Meteor scale being proportional to PE effort only up to scores of approximately 0.6.

Another potential explanation for this phenomenon may be related to the fact that evaluation metrics based on textual similarity are known to lose sensitivity for higher-quality MT (Babych and Hartley 2008), which could have led Meteor to overestimate the quality of high-scoring sentences. Babych and Hartley highlight that automatic evaluation metrics are more suitable for capturing problems at a lexical level, while a human assessment or a task-based evaluation (e.g. PE) would also be expected to indicate functional problems, including those pertaining to the text as a whole. Indeed, to a greater or lesser extent, this should apply to any MT evaluation metric based on word similarity, including Meteor. Even though participants did not have access to the entire text at once in the eye-tracking task (which could have prevented document-level issues from being attended to), both eye-tracking and think-aloud results show a curved relationship between Meteor and effort indicators. This could be due to the lack of sensitivity of MT evaluation metrics reported by Babych and Hartley.

5.1.1.2. *Source-Sentence Length*

Regarding sentence length, longer source sentences were perceived by participants as more cognitively demanding, as denoted by results in the mixed-effects model predicting subjective cognitive effort (see section 4.1.2). In addition, in the model predicting average fixation duration (a normalised measure not expected to vary with sentence length alone), Meteor was found to significantly interact with the length of source sentences, having a stronger negative correlation with average fixation duration for longer sentences. In editor-behaviour terms, this suggests that if two MT sentences have a high Meteor score and their corresponding source sentences have different lengths, lower cognitive effort would be associated with the longer source sentence. It

[66] The reader is reminded of the Meteor scale, which ranges between 0 (no similarity with reference translation) and 1 (perfect match with the reference).

should be noted, however, that a bias of automatic metrics towards sentence length is a known issue in MT evaluation. The BLEU score (Papineni et al. 2002) and its variations are known to favour short sentences (Nakov, Guzmán and Vogel 2012). By contrast, in the dataset used by O'Brien (2011), sentences that fell into the low GTM band (i.e. poor quality) were found to be substantially shorter than the remainder of the sample. Meteor seems to follow the same tendency as GTM, *penalising* shorter sentences. As the length of MT sentences is expected to mirror the length of their source counterparts, the interaction between Meteor and source sentence length observed for average fixation duration seems to be a potential artefact of this bias, rendering the score more reliable for longer sentences. This seems a more plausible explanation for this effect than participants having a true behavioural tendency to expend less effort on a longer sentence relative to a shorter sentence of the same level of MT quality.

It might also be that the very nature of the average fixation duration measure is able to explain the interaction between Meteor and sentence length. A longer sentence requires a larger total number of fixations. This could perhaps shorten individual fixations (and the overall average) as a result of more stimuli being available, i.e. more content to look at. However, in view of the known bias of MT automatic evaluation metrics towards sentence length, an artefact of eye-tracking metrics in this respect seems unlikely, or at least a minor factor.

5.1.1.3. *Source-Text Type-Token Ratio, and Prepositional and Verb Phrases*

As mentioned by Aziz, Koponen and Specia (2014), the PE effort associated with ST features may be explained by two converging phenomena: the fact that some of these features are difficult to process by their own nature, even in a monolingual context, and the fact that some of these features are problematic for MT, resulting in translations of lower quality.

It was observed in section 4.1.2 that word repetition in the source sentences (in terms of type-token ratio) was associated with high levels of cognitive effort. This effect is more likely to act as a mediator of translation quality. Even though type-token-ratio is not frequently mentioned in the literature as a negative translatability indicator (i.e. a ST feature that poses problems for MT), it could be hypothesised that post-editors find it difficult to deal with repeated words close to each other not because this is in itself a linguistic feature that is mentally hard to decode, but because repetition in the ST may account for a lack of flow in the MT output – e.g. the French string *Je recommande donc le test à partir de 50 ans, ou à partir de 40 ans* ['I therefore recommend the test from the age of 50, or from the age of 40'], whose machine translation was 'I recommend the test therefore to leave from 50 years, or to leave from 40 years' (see Ex.10, p. 104). While this case constitutes the repetition of errors (e.g. 'to leave from') that would need to be corrected even if there was no repetition in the sentence, word repetition could be seen at least as an aggravating feature in this sentence.

An unnatural amount of function words in the MT sentence could also be an explanation for the negative effect observed for type-token ratio. It was noticed from the materials that MT sentences corresponding to source sentences with low type-token-ratio (i.e. more word repetition) tended to have repetitive articles, pronouns and propositions that hindered the flow of the sentence. For example, the repetition of the definite article in the MT string 'The new restrictions allocate the youths, the minorities, and the people' (see Ex.16, p. 124), which was edited by most participants.

As for the relationship between prepositional phrases and cognitive effort, this effect could perhaps be explained as both a translation quality mediator and as a direct trigger of higher PE difficulty. Prepositional phrases are cited in the machine translatability literature as a challenging textual characteristic to be handled by MT systems (Underwood and Jongejan 2001). At the same time, this feature reflects syntactic embedding, e.g. with phrases being connected to each other by the preposition 'of'. A higher incidence of embedding of this kind may require more working memory capacity for connections between the head term or phrase and embedded chunks to be mentally kept on hold whilst decoding the sentence. The MT string '180 draft laws restricting the exercise of the right to vote in 41 states were introduced', presented in Ex.11 (p. 105) and Ex.17 (p. 123), seems to be a good example of this effect. Even in a monolingual context, the cascade of prepositional phrases in this sentence may hinder an immediate realisation that verb and object (i.e. 'introduced' and 'laws') are on the opposite ends of the sentence. It is hypothesised here that in addition to being a problem for MT, from a cognitive perspective, this may also pose difficulty to post-editors.

Regarding the effect observed for verb phrases, since verbs are the central elements of a clause, a high incidence of verb phrases may indicate the amount of information conveyed by the text (similarly to lexical density; see section 3.4.1.1). This could be an explanation for the positive relationship found here between verb phrases and the levels of cognitive effort participants reported to expend, with verb phrases being likely to act as a direct source of difficulty as opposed to an MT-quality mediator. However, the correlation observed here between cognitive effort and verb phrases was relatively weak, so further research would be necessary to confirm the effect of this feature.

5.1.1.4. *Estimating Cognitive Effort before Post-Editing*

As sentence-level predictors of cognitive effort, it is fair to say that MT evaluation scores that rely on human reference translations, such as Meteor, are relatively limited from the perspective of post-editors, as PE would not be necessary if a human reference translation was already available. However, it is worth noting that results obtained here are able to stimulate further research on (reference-free) quality estimation metrics (Specia 2011) (see section 2.3.2.2), which aim at predicting MT quality based on textual features that correlate with PE effort. This is still an emerging area that involves a high level of computational complexity, so the impact of quality estimation scores on PE effort could not be tested in this thesis. Nevertheless, these scores rely on some of the textual characteristics examined in the present study.

As mentioned above, results obtained here suggest that the type-token ratio and frequency of prepositional and verb phrases in the ST function as good predictors of cognitive effort in PE. Since these relationships were observed over and above any effects of MT quality (as measured by Meteor) and post-editors' individual characteristics, this gives good indications that these ST linguistic features should be further explored or perhaps given more weight in the context of quality estimation for the French-English language pair.

MT evaluation metrics based on textual similarity (such as Meteor) are also used as predictors in quality-estimation systems. This is done by exploiting translations produced by MT engines other than the one being evaluated as 'pseudo' references (see e.g. Specia and Shah 2013). In other words, for a given MT sentence whose quality needs to be predicted, Meteor scores would be computed based on the similarity between the sentence to be assessed and other MT versions of the same corresponding source sentence, which may not necessarily be correct but can nevertheless act as a parameter for comparison (hence the name 'pseudo' reference). Based on the design adopted here it can be concluded that Meteor is a generally good predictor of cognitive effort in PE. However, it was observed that this metric is a less reliable predictor in the context of extremely high scores and a more reliable one when used to assess longer sentences. This could be implying that if quality estimation is attempted for a specific genre with particularly short sentences, Meteor would be less informative as a predictor of PE cognitive effort and, by extension, of raw MT quality. Further research into quality estimation could be carried out in this respect to inspect if Meteor scores would behave in the same way when computed based on ‘pseudo’ reference translations. This could lead to further experimentation regarding different ways of using Meteor depending on its contrast with other textual features.

The relationship between ST linguistic features and MT evaluation metrics, rarely explored in previous research, seems of particular relevance in predicting MT quality and PE effort. In the present study, Meteor was more strongly associated with cognitive effort than ST features. Assuming Meteor as an indicator of MT quality, this denotes that, all things considered, the quality of the MT output has a stronger relationship with cognitive effort than characteristics of the ST. The ST is of course also important since some of its features may relate to PE effort in ways not accounted for by MT quality alone. However, since certain features of the ST are known to be linked to problematic MT outputs, the impact of significant ST predictors observed in previous research may well be more strongly predicted by MT quality measures when these are also taken into account in the analysis. This reinforces the idea that research in PE should shift to more integrated strategies accounting both for MT quality and for ST characteristics, as investigating either of these two elements in isolation arguably constitutes an incomplete approach to the problem.

5.1.2. *Post-Editors’ Individual Characteristics and Cognitive Effort (RQ2)*

The second research question (RQ2) addressed in this thesis concerns the relationship between post-editors’ individual characteristics and the amount of cognitive effort they expend during the task. This question could be regarded as an extension of RQ1 (i.e. the impact of *textual* characteristics on cognitive effort), since the impact of participants’ individual characteristics should also be controlled for when examining correlations between textual features and cognitive effort. In view of this, RQ2 constitutes both a question in its own right and an aspect that needs to be addressed so RQ1 can be properly handled. Findings in this respect point to complex interconnections. Three tendencies were observed regarding the relationship between cognitive effort and post-editors’ individual traits: (1) post-editors with more professional experience (in either translation, traditional revision or PE) reported expending higher levels of cognitive effort than their counterparts when post-editing MT outputs with higher Meteor scores (i.e. higher quality); (2) post-editors with a higher level of proficiency in the source language also reported expending more cognitive effort when post-editing sentences with higher Meteor values; and (3) in the case of post-editors with lower levels of proficiency in the source language, a higher amount of ST consultation was found to be associated with higher levels of subjective cognitive effort.

As for tendencies (1) and (2), these findings suggest that a higher level of source-language knowledge and professional experience does not necessarily alleviate the expenditure of cognitive effort in the activity. More source-language knowledge may allow post-editors to identify problems in the MT output that would otherwise go unnoticed (and therefore pose no effort). The same effect may be triggered by high levels of professional experience, which could account for a more acute perception of subtler issues in the text. Here, previous research suggests that experts distribute effort differently from novices in translation by identifying more problems and by being able to solve recurrent issues automatically whilst allocating cognitive resources to problems that require more control (Jääskeläinen 2010, 221). This may also be the case in PE, a hypothesis that seems to act as further support for tendency (1).

Tendency (3) indicates that, unsurprisingly, consulting the ST is more effortful for participants with little knowledge of the source language. Here, previous research shows that PE can lead to improvements in quality even when participants have no access to the ST (i.e. in ‘blind’ PE), which applies especially to fluency (i.e. linguistic quality as opposed to translation accuracy) (Mitchell, Roturier and O’Brien 2013). Similarly, results obtained in this thesis point to only slight differences in post-edited fluency between participants with high and low levels of proficiency in French (see section 4.1.2.1). On the other hand, source-language knowledge was found to have a significant effect on the adequacy of the post-edited text, as results in Table 12 (p. 117) indicate. These findings suggest that if PE is at all to be carried out by individuals with little source-language knowledge, impeding access to the ST might incur lower levels of cognitive effort with little effect on post-edited fluency, but with an imminent risk of adequacy errors. In situations where the post-edited text is short-lived and not intended for publication, however, appointing monolingual post-editors may well be a realistic practice, as adequacy errors would be expected to have less serious consequences in a context of this kind.

For those with a high level of source-language proficiency, the act of consulting the ST did not in its own right correlate with higher levels of cognitive effort. In this respect, the fact that participants with higher levels of source-language proficiency expended more effort when post-editing sentences with higher Meteor scores suggests that effortless ST consultations may have allowed more cognitive effort to be focused on the MT output in the case of these participants. It might also be that, in the context of high-quality MT, these participants spent more time and effort looking for potentially minor non-equivalences between the MT output and the ST. This was clear from the think-aloud task. P25, for example, a participant with a high level of French (98/100), declared on one occasion after reading a good-quality MT sentence: ‘that sounds good, but I don’t know if it’s the same in French’. This illustrates that the larger amounts of effort expended by participants with higher levels of French when post-editing good-quality MT could stem from a more thorough approach to the activity, with these participants checking that all information in the MT output corresponds to the source. In view of the high level of French proficiency of these participants, however, this would not be expected to require lengthy interpretations of the ST but rather a lengthy evaluation of the MT output itself.

Regarding other characteristics of participants themselves, no significant effects of working memory capacity were observed. Even though previous research in monolingual text revision has found a positive impact of working memory capacity on task performance (Francis and McCutchen 1994, in McCutchen 1996), results in this respect are mostly qualitative, with the authors stressing that further research would be required to examine the generalizability of the findings. Based on the insignificant results obtained here, it can be assumed that working memory capacity is at best a minor factor in PE. While a descriptive (i.e. non-inferential and therefore non-generalizable) effect of working memory capacity on PE productivity was found by Vieira (2014) based on group comparisons, it seems that this effect is harder to capture in the more rigorous context of a multiple regression analysis (see section 3.4.5.1), where other potential factors, such as MT quality and participants’ source-language knowledge, are also considered. This is likely to be the reason why positive effects of working memory capacity found in related areas (see e.g. McCutchen 1996) could not be observed here for PE.

More generally, cognitive effort is essentially a subjective construct in that it reflects how *individuals* respond to the demands of a task (see section 2.2.4). In view of this, impacts of subjects’ individual characteristics on the measures of cognitive effort exploited in this thesis would be expected. However, the fact that stronger effects of this kind were not observed does not seem uncommon, with previous research reporting similar findings. De Almeida (2013), for example, found no correlation between PE time and post-editors’ level of professional experience, suggesting that these patterns are too complex to be captured by temporal values alone (see section 1.1). In the present study, based on measures of cognitive effort, correlations of this kind were not observed either, so rather than an insensitivity of temporal measures to effects of this kind it seems that the complex nature of the task coupled with large variations in participants' behaviour may be a more plausible reason why the relationship between subjects' individual characteristics and PE effort is hard to capture.

Interestingly, however, the TAP-based analysis carried out in section 4.2.4 showed that post-editors who had more professional experience produced a larger number of TAP units coded with categories representing specific linguistic aspects of the task, such as grammar, lexis, discourse, knowledge and orthography. This suggests that the processes employed by these participants have more specific goals, denoting in any case a more explicit approach to the linguistic issues dealt with during the task. Those with a more positive attitude towards MT produced fewer TAP units coded with the Style category, which suggests they were less critical about the style of the MT output or the TT emerging from their editing. These findings once again go to show that complex relationships of this kind abound in PE, with the profile of a good post-editor being an apparent open issue in the field. These results also suggest that while the impact of participants' individual characteristics on PE behaviour is hard to predict in terms of cognitive effort, this impact is more evident when the *nature* of participants' cognitive processes is taken into account. This is not central to the present thesis, so a qualitative account of cognitive processes as a function of post-editors' individual characteristics could be further explored in future research, ideally by drawing a parallel with existing literature in this respect available for translation.

5.1.3. *Cognitive Effort and What Post-Editors Actually Do (RQ3-4)*

Differently from RQ1, which is based on information derived from the ST and the MT output, RQ3 and RQ4 concern aspects deriving from the process of carrying out the activity, namely the nature and frequency of different linguistic aspects of the task post-editors attend to (RQ3) and a potential association between these aspects and cognitive effort (RQ4). Here, the PE process was found to be in most part centred on reading, grammar and lexis, with grammar and lexis also being significantly associated with the amounts of cognitive effort expended by participants.

5.1.3.1. *Lexis*

Regarding the high incidence of TAP units coded with the Lexis category, it is postulated here that this pattern may generalise to most kinds of translating activities, as mentioned in section 4.2.2.3. In this respect, it is interesting to note that Lexis TAP units did not concern only outright accuracy problems in the transfer between source and target. These units were also related to the suitability of different terms in the context of the target language (e.g. collocation issues such as ‘surveilling a disease’ instead of ‘monitoring a disease’). Indeed, previous research shows that yielding outputs with correct collocations can be particularly challenging for MT systems (Vieira 2012). Here, intuition suggests that lexical issues are more target-language-centred in PE than in translation, with post-editors solving a potentially larger number of fluency problems (i.e. accurate but disfluent lexical choices) as well as solving most accuracy problems by stumbling upon the inadequacy of terms in the MT output as opposed to having to mentally search for lexical alternatives directly from the source. This is of course related to the fact that an existing target version is available in PE (i.e. the MT output), which means that the role played by the target language may also be more central in the case of traditional translation revision.

In the context of translation self-revision processes, Shih (2006, 175) highlights that the prominence of lexis may be explained by the fact that lexical items are amongst the shortest and most basic units in the language, being therefore readily available for modification. While this explanation may in part also apply to PE, data obtained here suggests that, despite involving small linguistic units, lexical issues can also be cognitively demanding. In native-language writing, for example, it is generally acknowledged that lexical retrieval (i.e. finding adequate lexis in long-term memory) can be extremely effortful, with lexical writing processes putting a strain on the quality of attention devoted to other aspects of writing tasks (see McCutchen 1996; Schoonen, Snellings, Stevenson and Gelderen 2009, 90). It could be argued that the process of substituting lexical items in the MT output is not dissimilar, as post-editors are also required to search long-term memory when attempting to find different lexical alternatives.

Indeed, solutions for problems that concerned mostly the target language did not seem to come easily to participants, as in the example involving the string ‘surveillance of the disease’ (Ex.20, p. 132), where most participants failed to replace ‘surveillance’ with ‘monitoring’, a more appropriate term to be used with ‘disease’. In this specific case, it seems that exploiting syntagmatic relations in the sentence (i.e. how words are linked to each other) is a less demanding and more effective strategy in solving problems of this kind. P24 solved this problem the first time she came across it, declaring ‘I think you probably ***monitor*** [~~*surveillance*~~] a disease’. It is clear from this statement that P24 attempted to exploit the link between ‘surveillance’ and ‘disease’ as opposed to just trying to think of a synonym for ‘surveillance’, a strategy employed with no success by P25, for example, who declared ‘I’m not very keen on *surveillance*, but I can’t think of a better word’. It was only when P25 saw the verb ‘to monitor’ elsewhere in the text that she realised that ‘monitoring’ was the word she was looking for. This is consistent with psycholinguistic research in word processing, which indicates that establishing mental links between words that frequently co-occur in the language (e.g. ‘monitoring’ and ‘disease’) is easier than establishing links between words that are simply semantically related (e.g. ‘monitoring’ and ‘surveilling’) (Traxler 2012, 86-87).

It seems then that an over-use of purely semantic connections, on a paradigmatic axis, could be one of the reasons behind participants’ struggle in trying to find appropriate synonyms to replace inadequate words in the MT output. It is noteworthy in this respect that, even in a context where stylistic changes are not recommended, recent research suggests that MT mistranslation errors, including those involving false cognates and terminology, are correlated with average pause ratios in PE (Lacruz, Denkowski and Lavie 2014). Findings in this thesis follow this trend, but, instead of considering just MT errors, all aspects of PE are taken into account here, including edits that are not actually implemented but only mentally planned.

5.1.3.2. *Grammar/Syntax*

While a strong relationship was observed here between cognitive effort and the frequency of units coded with the Grammar category, certain grammatical problems seemed easier to handle than others. The problem illustrated in Ex.18 (p. 131), for instance, involving the adjective ‘important’ being used as a noun in the string ‘the important is to…’, prompted a quick reaction from most participants, who seemed to effortlessly add the word ‘thing’ (i.e. ‘important *thing*’) to the MT output. Other issues seemed more demanding, such as unclear pronoun referents – e.g. the ambiguous ‘it’ in the string ‘it will suggest to active monitoring’ (Ex.19, p. 132). This apparent variation in the amount of effort posed by grammatical issues seems largely explainable by the reasoning justifying the MT error typology proposed by Temnikova (2010), i.e. that errors that stretch across longer textual spans are generally harder to deal with. In the case of ‘the important’, for example, no long-span searches in the text are necessary to realise that the phrase requires the noun ‘thing’; it suffices to look at the next word. As for ambiguous pronouns, post-editors need to look for the correct referent elsewhere in the text whilst keeping the problematic string in mind, which can put more pressure on working memory, causing more effort.

Even though the issues mentioned above seemed to require different levels of effort from participants in general, it is worth pointing out that both these issues occurred in sentences that fell into the high-cognitive-effort cluster. This suggests that measuring cognitive effort at a sub-sentence level seems like an interesting direction for future research. While Aziz, Koponen and Specia (2014) analyse the PE process at a sub-sentence level based on the edits post-editors implement in the MT output, using traditional cognitive-effort measures for this purpose would be more troublesome. Gathering sub-sentence subjective ratings, for example, would require participants to engage in extremely repetitive behaviour while the use of eye tracking would require gaze-to-word-mapping algorithms which, as mentioned in section 3.3.4, are still relatively fuzzy. This suggests that there is considerable room for further methodological developments in translation process research that would allow cognitive effort and other complex constructs to be estimated at higher levels of precision. As seen in section 4.3, TAPs would also be an alternative in this respect, which is discussed in detail in section 5.3.

5.1.3.3. *Grammar vs. Lexis*

All in all, Grammar and Lexis were the coding categories that presented the strongest relationship with cognitive effort based on results obtained in the think-aloud task. In Temnikova's (2010) MT error typology, 'incorrect style synonym' and 'incorrect word' (i.e. lexical issues) are ranked at levels 2 and 3 of cognitive demand, with 10 being the most demanding level. This typology was empirically tested by Koponen et al. (2012) in view of PE time. While Koponen et al. propose a few modifications to the ranking of errors on the list (see section 2.4.2), lexical errors were found to be associated with less PE time than syntactical ones. Results obtained here, which concern the task in general (including issues pertaining to the ST), mirror this trend, i.e. with cognitive effort having a stronger association with grammar/syntax than with lexis. This difference, however, was not found here to be significant, a possibility that is not examined by Koponen et al. or Temnikova. While in a large number of cases processing span (i.e. how far apart errors stretch in the text) seems indeed to be a good explanation for the amount of cognitive effort expended by participants – e.g. in the context of grammar/syntax issues – it appears that processing span alone is not able to explain all the cognitively demanding scenarios that PE entails. In the case of lexis, for instance, processing span would not be expected to be a factor, so it is suggested here that the demands of lexis should not be underestimated in PE and that lexical and grammatical aspects of the activity should receive equal attention, be it, for example, in post-editor training or in effort estimation.

5.1.3.4. *Other TAP Categories*

Regarding other TAP categories, a complex relationship was observed between Style units and cognitive effort while no relationship in this respect was noticed in the case of Orthography, Knowledge or Discourse units. Here it should be noted that the fact that these categories were relatively infrequent could potentially account for these insignificant results. This, however, is not regarded here as problematic. The TAPs dataset has over five thousand units overall, so the relative infrequency of these categories seems to be a characteristic of PE itself rather than an artefact of a small quantity of data being available. In other words, a bigger dataset, based on the same genre and domain, would be expected to have a similar distribution of TAP categories. Nevertheless, larger absolute TAP unit counts would give more statistical power to the analysis, which could potentially reveal subtler effects. A larger absolute number of data points might also be able to confirm or challenge the complex relationship observed between style and cognitive effort, found to be relatively weak.

Stronger links with different aspects of the task may be the reason why categories such as Knowledge and Discourse were not found to have a significant effect. Knowledge would be expected to be highly dependent on the nature of the subject matter and genre of the text. Since neutral texts that were relatively undemanding on subjects' world knowledge were chosen here, it may be that the infrequence of Knowledge TAP units is in part a result of this choice. Attending to aspects of discourse, in turn, was an attitude more prominently observed in the second-pass phase of the task, i.e. in the case of participants who checked their work at the end. In view of this, it seems that the Discourse coding category is more clearly associated with a specific stage of the task rather than with the levels of cognitive effort expended by participants. As for the Orthography category, TAP units involving spelling corrections were associated mainly with MT output of low quality, as would be expected. Post-editors are instructed to correct spelling by traditional PE guidelines (TAUS/CNGL 2010), but corrections of this kind would of course only be expected to occur in contexts involving poor MT.

5.2. Post-Editing Effort and Post-Edited Quality (RQ5)

The fifth research question (RQ5) addressed in this thesis concerns the impact of cognitive and other types of effort on post-edited quality. Findings in this respect denote that editing interventions function as a better predictor of final quality than the amount of eye fixations landing on the text, found to have a negative association with both post-edited fluency and adequacy. Even in the case of editing interventions, however, it was found that changing the MT output has a negative effect in the context of high-quality MT. This is in line with professional guidelines that advocate the implementation of as few changes as necessary in any PE scenario. Naturally, these guidelines would be expected to entail higher levels of productivity, but, perhaps surprisingly, it is shown here that an economical editing approach may also entail higher levels of post-edited quality. Potential reasons explaining this and other aspects of the interplay between effort and product quality in PE are discussed below.

5.2.1. *A Seamless Integration between Man and Machine*

It was observed in the present study that one of the reasons behind post-edited sentences of lower quality might be an incompatibility between the baseline machine suggestion and some of the post-editors' modifications. It seems post-editors may at times envisage their own preferred way of conveying the ST message, which may come into conflict with the raw MT output, requiring a more radical editing approach to form a coherent edited text. In Ex.15 (p. 120), for instance, P13 regarded the string 'contamination between groups', in the MT output, as problematic, opting instead for 'data contamination between groups'. P13's initial modification of this string triggered edits outside the original text segment being edited, resulting in a decrease of quality. Indeed, the original machine suggestion 'contamination between groups' seemed to work well in this case, as denoted by sentences post-edited by other participants who did not alter this part of the MT output and achieved high fluency scores on this sentence. It seems then that further editing would be required for P13's suggestions to work well, which does not seem like a viable alternative when simply maintaining the MT output in its original state may well prove to be equally effective.

When working towards a coherent text that comprises both untouched raw MT and edited passages, especially in situations where time pressure may be a factor,[67] it seems that post-editors run the risk of prematurely abandoning the editing process for

[67] Though a blanket time limit has not been imposed in the present study, participants were instructed to carry out the task as fast as possible.

certain text strings when they engage in radical editing of the kind described above. For instance, when post-editing the machine-translated sentence ‘The new restrictions allocate the youths, the minorities, and the people in a disproportionate way to weak income’ (see Ex.16, p. 124), P17 used the word ‘groups’ twice, resulting in the string ‘young persons, minority groups, and low-income groups’. Two judges rated P17’s post-edited version of this sentence with 3 out of 4 on fluency. This score, which failed to achieve the highest level, is most likely connected with the repetition of ‘groups’, which was either not detected or not deemed a problem by P17. It seems that failing to shift attention between shorter and longer textual spans leads to unresolved issues of this kind being left behind, which could be due to the difficulty posed by these local-global shifts, as mentioned, for example, by Mossop (2007, 38). In this specific case, at least the second occurrence of the word ‘groups’ could have been avoided if edits that were closer to the MT output’s original structure had been implemented, as observed in versions proposed by other participants who simply chose ‘people on low incomes’ as a solution for the problematic construction ‘people to weak income’, in the raw MT output.

In view of the detrimental effect of having to shift attention between short and long textual spans, it could be postulated that, from post-editors’ perspective, an ideal scenario would be correcting errors that only entail local edits, under low-cognitive-effort conditions (i.e. apart from when dealing with perfect MT where no edits are needed). While of course this depends on the quality of the raw MT output, results obtained here as well as in previous research (e.g. de Almeida and O’Brien 2010) suggest that post-editors’ own preferences may also influence the extent of edits implemented in the text, which can at times cascade into unnecessary levels of cognitive effort.

5.2.2. *Negative Links between Post-Editing Effort and Post-Edited Quality*

It was observed in section 4.1.3.2 that, for low- to medium-quality MT, increasing amounts of edits were linked to increasing post-edited quality. A different trend was observed for high-quality MT, where increasing amounts of editing, rather than being indifferent, were linked to *decreasing* post-edited quality. Fixation count was found to be generally negatively connected with the levels of quality of the post-edited text, which was a particularly startling finding. As fixation count was found to be correlated with PE time (see section 4.1.3.2), the results just described suggest that, for low- to medium-quality MT, more edits in less time have a positive impact on quality. This finding is quite surprising as it would arguably be expected that the more participants fixate on the text (either the ST or MT-TT) the better their post-edited output would be. However, a number of reasons seem to counter a positive and linear relationship between fixation count and post-edited quality. Some of these reasons could be related to the methodological design adopted in the study, while others seem intrinsically rooted in PE. Explanations falling into these two groups are discussed separately, below.

5.2.2.1. *Potential Artefacts of Study Design*

Regarding the negative relationship observed here between fixation count and post-edited quality, it should first be noted that participants in the present study were not allowed any 'drawer-time' – i.e. they could not distance themselves from the text and then look at it again with a fresh pair of eyes. Various drafts with periods of drawer-time in between is perhaps a context where an overall positive relationship may be observed between PE time and the quality of the post-edited text as participants would be more likely to attend to any unsolved issues left behind by working on various drafts. This would nevertheless be a time-consuming approach to the task, which seems inconsistent with the urgency normally associated with the post-edited product. PE is normally advertised as a 'quick solution', which applies even to situations where human or near-human quality levels are expected,[68] so it seems that the restriction on drawer-time imposed here is not unrealistic. A timely approach to the task may in fact avoid cascades of further problems that would require several drafts to be resolved (see Ex.15, p. 120).

It should also be taken into account that participants were not allowed to backtrack in the eye-tracking task. In a setting of this kind all problems in a given sentence need to be solved 'in one go', a situation where more time spent looking at the sentence, without the distancing effect mentioned above, could have induced the expenditure of higher levels of redundant effort, i.e. effort that is not reflected in final quality. However, this alone does not seem able to explain the negative relationship observed between fixation count and post-edited quality. In the think-aloud task, where backtracking was allowed, PE time (which correlates with fixation count) did not necessarily have a positive impact on quality either, with subjects such as P27, who did not have professional experience in translation or PE, finishing the task in relatively little time without an obvious detrimental effect on

[68] See e.g. http://www.lingo24.com/pemt.html (Accessed July 2015).

product quality being noticed in comparison with other participants who took longer to finish the task (see section 4.2.1). While a text-level quality assessment of the post-edited texts produced in the think-aloud task would be able to shed further light on these connections, TAP unit counts were found to produce similar results in comparison with measures of cognitive effort gathered in the eye-tracking task, suggesting that the impact of the backtracking restriction is not significantly large.

Regarding the detrimental effect of many edits observed for high-quality MT, it cannot be excluded that this result could be in part a consequence of the categorical scale adopted in the study to measure quality. As this scale imposes an artificial limit to the amount of quality improvement that can be achieved through PE, i.e. up to level 4 (out of 4), the best case scenario concerning raw MT sentences already deemed to be at level 4 is remaining at this same level after PE. In other words, post-edited quality cannot carry on increasing ad infinitum based on a categorical scale. This could have meant that any random error in the assessment could have made the trend go downwards in the case of MT sentences at level 4. This explanation is nevertheless not regarded here as the most plausible one, which is grounded in two reasons. First, there was a clear connection between amount of changes and the decrease in quality observed for raw MT sentences at level 4; when few or no changes were implemented, post-edited sentences remained at level 4, suggesting that it would be unlikely for the decrease in quality observed to be an effect of error or inconsistency in the assessment. Second, this trend is consistent with professional guidelines and with previous research where the pitfalls of over-editing are highlighted (Guzmán 2007). To the present author’s knowledge, this thesis is the only PE study to date where the dangers of too many edits are empirically observed whilst testing for the impact of different types of effort – e.g. cognitive and temporal (Krings 2001) – and whilst also taking different levels of (raw) MT quality into account. Directions for further research in this regard are proposed in the next chapter.

5.2.2.2. *Potential Benefits of a Fast Editing Behaviour and the Role of Intuition*

Interestingly, a negative association between PE time and post-edited quality has been observed in recent research, similarly to results on fixation count obtained here. Green et al. (2014) were interested in examining the feasibility of using interactive MT systems in PE, i.e. when the MT output adapts itself to post-editors’ edits on the fly. In doing so, they found that PE time was a negative predictor of the quality of the post-edited text, measured at a sentence level with the automatic metric BLEU+1. The authors explain this phenomenon by highlighting that more PE time will inevitably be spent post-editing longer sentences, which are known to be penalised by the BLEU score (see section 2.3.2.1). However, the correlation between PE time and source-sentence length the authors report (ρ = 0.53 for French-English and ρ = 0.43 for English-German) is relatively moderate, which could be indicating that there may be more to this negative effect than just the fact that the BLEU score tends to assess long sentences as being of lower quality.

As results obtained in this thesis are based on a human assessment of the post-edited text, with fixation count being normalised by source-sentence length and with potential effects of post-editors' profiles being controlled for in the context of multiple regression, it seems that a negative association between post-edited quality and fixation count (or PE time) could be a phenomenon intrinsic to PE.

It was observed in section 4.1.3.2 that a dense editing behaviour (i.e. more edits in less time) seemed to have a positive association with the fluency of the post-edited text when the MT output was at low or medium levels of quality. It could be hypothesised that this effect is related to the benefits of having fast reactions to the problems stumbled upon in the text, which could avoid an overly analytical approach to the task. Offersgaard, Povlsen, Almsten and Maegaard (2008) suggest, for example, that being able to make quick decisions on machine-translated strings that can be retained and those that should be discarded is key to successful PE performance. This raises the question of what accounts for the efficacy or inefficacy of these decisions.

Researchers in decision making normally assume the existence of a 'dual model of thought' (Hogarth 2012, 67), whereby decisions are made as a result of both tacit and deliberate systems, i.e. by intuition and as a result of deliberate (controlled) analytical thought, respectively. This dual mode of thought is also related to the notion of automaticity in translation, i.e. the fact that certain translation issues are dealt with automatically while others require a controlled approach (see e.g. Jääskeläinen and Tirkkonen-Condit 1991; Tirkkonen-Condit 2005) – see also section 2.2.2.3. To the knowledge of the present author, empirical studies on the efficacy of each of these operating modes (automatic vs. controlled) have not been carried out in PE to date, though Martín (2010, 182) suggests that both these modes are important factors in translation and that the role of ‘unconscious’ (i.e. automatic) processes should not be underestimated.

In a general context, Hogarth (2012) hypothesises on the different elements that may influence the efficacy of intuitive and deliberate decision making. He proposes a chart predicting that for tasks of moderate to high levels of complexity, intuitive decisions are expected to outperform analytical ones provided that the amount of error induced by automatic processes is kept at low or medium levels (ibid., 77). In other words, if intuitive decisions are likely to be accurate and the task at hand is a complex one, acting by intuition may represent a better and more efficient approach.

Hogarth (2012) posits that intuitive decisions are likely to be accurate only when individuals have been previously exposed to favourable learning conditions, having had a chance to act upon constructive feedback (ibid., 76). All participants in the present study had had some experience with translation, either as students or professionals, so it is relatively safe to assume that they have to some extent experienced learning conditions of the kind described by Hogarth, even if not specifically in the context of PE. It seems then that intuitive textual modifications would carry low risk in the case of these participants, or perhaps medium risk in the case of those with lower proficiency in French. As it is also fairly safe to assume that PE is a complex activity,[69] it is hypothesised here that the positive effect of a dense editing behaviour could be at least partly explained by the efficacy of intuitive decision making. Particularly in the case of expert participants, it seems that making decisions that are more intuitive and less analytical could prove capable of lessening the dangers of ‘over-editing’. Intuitive editing would also be less cognitively expensive, with decisions relying directly on long-term memory as opposed to overloading the limited capacity of the working memory system (see section 2.2.2.1).

A high incidence of automatic processes in the task would also be able to explain the generally negative association observed here between PE effort and post-edited quality. This is because automatic processes are not usually accounted for by measures of effort. Controlled processes are expected to take more time, being associated, for example, with more fixations and being available in working memory, which allows these processes to be taken into account by participants

[69] As affirmed, for example, in the case of translation (de Groot 2000).

when they provide subjective ratings and when they think-aloud during the task. Automatic processes, on the other hand, are faster and should have a less appreciable connection with eye movements or any other measure, though still leading to improvements in the MT output. In the context of subjective ratings, for example, it is generally assumed in previous research that a high incidence of automatic processes in the task may account for a disassociation between effort and performance (i.e. the quality of the results obtained) (Gopher and Donchin 1986, 24).

At this point it should be noted, however, that a clear-cut dichotomy between 'intuitive' and 'analytical' processes is likely to be an oversimplification of the matter. A safer assumption would be that editing processes could be placed upon a continuum of deliberate control, as suggested by Martín (2010, 180). Indeed, based on the methods adopted here, it cannot be guaranteed that a dense editing behaviour is *necessarily* linked to a larger amount of intuitive edits. Confirming this would require making use of methods capable of distinguishing between edits made by intuition and those that require analytical thought, such as the dual task strategy employed by Larigauderie, Gaonac'h and Lacroix (1998) – see section 2.2.2.1.

Ultimately, Hogarth (2012, 80) suggests that 'people should be more aware of how often they allow themselves to take decisions automatically'. Based on results obtained here, it seems this could be key to revealing further information on the links between effort and product quality in PE.

5.2.3. *Staying within the Window of Optimum Behaviour*

Seen widely, results obtained in the context of RQ5 (i.e. the relationship between PE effort and post-edited quality) suggest that the most efficient way of distributing effort in PE involves envisaging a point where edits are not too few and not too many, with situations that induce cognitive effort – e.g. cascades of edits that require constant shifts between short and long textual spans – being kept to a minimum. This ideal point could be thought of as an optimum window, i.e. the region within a spectrum of editing behaviour where cognitive effort would be as low as possible whilst keeping post-edited quality at high levels.

The relationship between these different factors is illustrated in Figure 24, which shows a spectrum of amount of editing that goes from a zero-edit situation (i.e. perfect MT) (left) to situations where large amounts of editing and potential retranslation are required (i.e. low-quality MT) (right). The optimum window of editing behaviour would start at a no-edit point involving little deliberation (i.e. perfect MT) and end at some point before retranslation, representing an operating mode where post-editors would carry out as few edits as possible, but as many as required to achieve an improvement of the raw MT output. While this is a straightforward reflection of the mechanisms of PE, based on results obtained in the present study it is possible to draw a parallel between these mechanisms and empirical information on cognitive effort and post-edited quality, aspects that received little attention in previous research. This is of course a tentative model, but findings obtained here in this respect suggest that operating within this window requires keeping levels of cognitive effort to a minimum, as it was found that more than being associated with lower levels of productivity, excessive cognitive effort may also entail lower levels of post-edited quality.

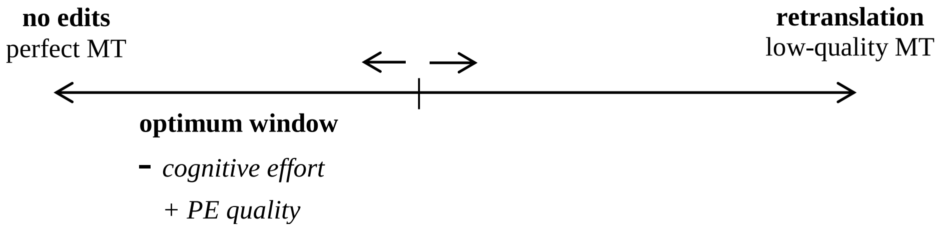


Figure 24 - Scheme illustrating window of maximum PE performance.

The exact factors underlying the ability to avoid unnecessary levels and/or types of effort are not entirely clear, but it seems that the fuzzy nature of certain translation problems may be behind the expenditure of redundant cognitive effort. In the context of translators' self-revision processes, Shih (2006, 180) observed that translators were better at solving problems that they were able to define. In the present study, it is interesting to note that raw MT sentences at fluency level 3 were the ones to present the largest number of cases where the MT output was not improved after PE (see section 4.1.3.1). This could be related to findings obtained by Shih: in view of the subtle nature of errors in MT sentences at fluency level 3 (e.g. the non-idiomatic use of 'rarer' as a modifier of 'number of people' in the string 'electoral fraud is rarer in the United States than the number of people killed by lightning' – see Ex.27, p. 147), it could be hypothesised that these sentences pose difficulty more in terms of defining what the problem is than in terms of actually correcting it. It could therefore be suggested that an ability to identify and define subtle problems in the MT output is central to operating within the optimum window illustrated in Figure 24. Furthermore, it seems that this might be better achieved through intuitive thinking, performed by trained individuals, as suggested by Hogarth (2012).

5.3. Think-Aloud Data as a Measure of Cognitive Effort in Post-Editing (RQ6)

To the knowledge of the present author, Krings (2001) conducted the only previous study exploiting TAPs as a measure of cognitive effort in PE. TAPs have come under criticism in previous years due to potential methodological problems involving lack of validity and completeness, as well as interference with the task (see e.g. Bernardini 2001; Jakobsen 2003; Li 2004). Similarly to the approach adopted by O’Brien (2005), TAPs were not assumed a priori as a measure of effort in the present study in view of these previous findings. Rather, TAPs were exploited here more as a window into the specific linguistic aspects of the task that were attended to by post-editors.

However, as suggested by Sun (2011), empirical evidence showing the *severity* of the interferences posed by the think-aloud condition is to date very limited. The sixth research question (RQ6) addressed in this thesis was aimed at shedding light on this issue by checking to see if TAPs correlate with more established measures of cognitive effort. Results in this respect suggest that, even though the think-aloud condition incurs more task time, the amount of TAP-based data produced in the task mirrors other measures of cognitive effort collected without any think-aloud interference.

In section 2.2.3.3, three potential weaknesses of TAPs (Krings 2001) were described in detail,[70] namely a potential disconnect between verbalisations and cognitive processes; the possibility that the think-aloud condition may *change* cognitive processes by interfering with the task; and the fact that TAPs may be an incomplete source of data. Findings obtained in the context of RQ6 are discussed below in view of these issues.

In regard to a potential disconnect between verbalisations and cognitive processes (e.g. because participants could be interpreting their actions as opposed to simply externalising them – see section 2.2.3.3), it seems that for the purpose of measuring cognitive effort this risk is no greater for TAPs than for other methods. Since TAPs, subjective ratings and eye movements were found to be correlated in the present study, it can be argued that these data sources are able to support each other’s claim to

[70] See also Shih (2006) and O’Brien (2005) for further discussion on these weaknesses.

measure similar constructs. In regard to eye tracking, it is worth remembering that the eye-mind and immediacy assumptions (Just and Carpenter 1980) – the rationale for the use of eye tracking as a method to investigate cognition – are not free from weaknesses, being criticised by theories that argue that eye-movement behaviour is driven by purely physical (i.e. non-cognitive) mechanisms. In this respect, it may be that the same assumptions made in previous research regarding the validity of eye-tracking data (see e.g. Hvelplund 2011) could be made here with regard to TAPs: while it cannot be guaranteed that all think-aloud verbalisations will reflect true cognitive processes, when the think-aloud method is properly deployed – e.g. with verbalisations being concurrent to the task (see Krings 2001) – this connection should exist in most cases.

As for the possibility of cognitive processes being *changed* by TAPs, previous research shows that the think-aloud condition incurs a larger amount of non-linear processes in PE (i.e. the mental consideration of more overlapping possibilities) (Krings 2001) and also a difference in how participants segment the text in translation, with shorter text segments being processed at a time (Jakobsen 2003). These effects were not investigated here in detail when comparing data from the think-aloud and the eye-tracking tasks as this has already been tackled by previous research. However, any potential effect of this kind did not prevent correlations between TAPs and other data sources from being observed. Indeed, it would arguably be expected that any differences in editing behaviour stemming from the think-aloud condition would remain relatively constant throughout the task, hence not impeding the observation of relative trends within a single dataset.

Concerning the potential problem that TAPs may be an incomplete data source (e.g. because participants may have different degrees of verbalisation willingness or because different thoughts can occupy working memory at the same time but cannot be verbalised simultaneously) it seems fair to assume that this problem is not exclusive to TAPs either, as argued by Krings (2001). Regarding eye tracking, for example, some participants may look at the keyboard more often while others may be able to touch-type, a situation that also leads to different amounts of data being produced by each participant. This is normally tackled in eye-tracking studies by establishing minimum thresholds that guarantee that enough data is available for analysis – see section 3.4.2.3. A similar approach could be adopted in the case of TAPs. In the context of this thesis, one participant was not retained in the study due to a difficulty in getting acquainted with the think-aloud condition, resulting in a very small amount of data being gathered. In this respect, it seems that the sub-field of translation process research suffers from a lack of standard approaches in dealing with data quality, which applies both to eye tracking and TAPs. In the case of the participant not retained in the analysis carried out here, hardly any data was produced and the participant declared finding it particularly difficult to think-aloud and carry out the task at the same time, which seemed like a good justification for exclusion. Nevertheless, decisions on what to retain and what to discard risk being arbitrary, which calls for further research aimed at defining standard practices of data-quality assurance in the specific context of translation studies.

All in all, results reported here indicate that, for the purpose of estimating cognitive effort, the interference posed by the think-aloud condition may be smaller than previously estimated, or at least not large enough to prevent correlations with other data sources from being observed. It seems that previous research looking into the methodological issues associated with TAPs tends to overlook some of the criticisms received by other methods. Perhaps central in this respect is the awareness that no method will be free from weaknesses, which seems to apply especially to empirical investigations on complex constructs such as cognitive effort, where the strengths of one method can compensate for the weaknesses of another. In this respect, while the mixture of cognitive-effort measures adopted here did not initially comprise TAPs, strong relationships observed in the study between different data types seem to confirm that TAPs too can be exploited as an indicator of cognitive effort in PE.

Chapter 6. Conclusion

6.1. Introduction

This thesis investigated the expenditure of cognitive effort in post-editing of machine translation. The concept of cognitive effort is regarded here as the amount of mental resources that individuals allocate to a task, which is distinguished from demands that are intrinsic to the task, in consonance with theories from educational and cognitive psychology. In line with previous research in PE and related areas, this construct was estimated through eye tracking (exploiting average fixation duration and fixation count as specific metrics) and a subjective scale of cognitive effort ranging between 1 (low) and 9 (high). Empirical data deriving from PE tasks, based on the French-English language pair, was then analysed in pursuing three specific aims:

Aim 1: Investigating the post-editing process with a view to identifying links between *cognitive effort* and elements pertaining to the *ST*, the *MT output*, *post-editors* and the *task itself*.

Aim 2: Investigating the relationship between (1) the levels of *cognitive and other types of effort* invested in the task and (2) the *quality of the results obtained*.

Aim 3: Investigating the relationship between *different types of data* that may be used as *measures of cognitive effort* in PE.

A number of research questions (RQs) were formulated in the context of these aims. Aim 1 comprises questions relating to how cognitive effort correlates with textual characteristics (RQ1), with characteristics of post-editors themselves (RQ2) and with different linguistic aspects of the task participants attend to (RQ4), whose frequency, nature and distribution are also investigated in their own right (RQ3). Aim 2 involves checking to see how PE time, cognitive effort and amount of editing are linked to the fluency and adequacy of the post-edited text (RQ5), while Aim 3 concerns mainly a methodological issue: whether or not think-aloud protocols (TAPs) correlate with eye movements and subjective ratings as measures of cognitive effort.

Two tasks were designed to answer these questions: one geared towards the use of eye movements and subjective ratings, referred to as the eye-tracking task, and a think-aloud task which provided insight into the mental underpinnings of the activity. These tasks were carried out by different, but comparable, samples of participants, with the same STs and MT outputs being used, which allowed a comparison of data produced under the think-aloud condition (which would inevitably interfere with any other method) with data from other sources. Conclusions based on results obtained in the context of these two tasks are described below in view of each specific aim of the study.

6.2. Cognitive Effort and the Post-Editing Process (Aim 1)

Information on a number of effort predictors was obtained in the context of Aim 1, which concerns how different elements of the PE process correlate with cognitive effort. The incidence of prepositional and verb phrases in the ST, ST type-token ratio, and Meteor, an automatic metric reflecting the quality of the MT output based on its similarity with a human reference translation, were all found to be good indicators of cognitive effort in PE. Information of this kind is particularly relevant for the purpose of developing systems capable of predicting effort. There are ongoing attempts in this respect in the field of MT quality estimation, an area that aims at providing MT quality scores that do not rely on human reference translations, but rather on textual features associated with PE effort. While it is generally accepted that *cognitive* effort is the main component controlling the overall expenditure of effort in PE (Krings 2001), to the knowledge of the present author cognitive effort has not to date been exploited in the context of MT quality estimation, which renders information on the effort predictors mentioned above extremely valuable for future initiatives in this respect.

Impacts of post-editors' individual characteristics on effort expenditure were also examined and it was generally observed that the relationship between textual characteristics and cognitive effort is much easier to capture than the relationship between cognitive effort and traits of post-editors themselves. This does not seem due to a lack of connection between post-editors' individual traits and the expenditure of effort in PE, but rather because complex links exist in this respect, as evidenced by interactions observed in mixed-effects regression models. In particular, consulting the ST was only found to be a sign of effort in the case of participants with a low level of proficiency in the source language, while post-editing high-quality MT output was perceived as more effortful by those with a *higher* level of proficiency in the source language. This goes to show that general statements assuming, for instance, that those with a higher level of professional experience will on all occasions expend less or more effort in PE are likely to oversimplify the complexity of these relationships. Effects of this kind seem to be moderated by other factors, such as MT quality.

Regarding the different linguistic aspects of the activity, it was found here that the vast majority of what goes on in PE involves dealing with lexical and grammatical issues or simply reading and evaluating the text (ST, MT output and/or emerging TT) without a specific problem-solving goal in mind. The levels of cognitive effort associated with lexis and grammar were found to be significantly higher than those associated with non-specific reading/evaluation and a number of lexical processes observed in the study were not centred on issues regarding the correspondence between source and target languages (French and English, respectively), but rather on target-language-specific substitutions (e.g. ‘talk’ vs. ‘debate’, or ‘monitoring’ vs. ‘surveillance’). This has a number of implications for PE practice as it gives indications of the potential benefits of making use of monolingual resources (e.g. thesauri) in the activity, an aspect that seems to merit further exploration in industrial settings as computer-assisted translation tools currently available rarely offer such resources as a built-in feature.

6.3. From Process to Product (Aim 2)

To the author’s knowledge, this thesis constitutes the first attempt at investigating how cognitive and other types of effort invested in PE correspond to the quality of the post-edited product, which was the second aim of the investigation. A number of surprising results were obtained in this respect. In particular, a negative relationship was observed between cognitive effort and post-edited quality overall. In addition, while the amount of changes implemented in the MT output was able to generally improve raw MT quality, this happened only up to a certain threshold, with the effect reversing in situations where raw MT quality was high.

Concerning the above effects, it was observed here that operating in a mode where changes are not too few and not too many (referred to in this thesis as a window of optimum behaviour) seems to involve the ability to identify and avoid situations that lead to the expenditure of excessive cognitive effort. This supports a minimal-editing approach to the task, which appears to have implications for post-editor training. In view of the results obtained in the context of Aim 2, PE students should arguably learn not only the linguistic aspects of how to edit MT output, but also how to operate according to different briefs, having the skills to distinguish between necessary and unnecessary edits and being able to identify situations that are likely to lead to excessive cognitive effort.

In regard to the relationship between post-editors' individual traits and the quality of the post-edited text, an expected positive impact of source-language knowledge on post-edited adequacy was observed, but, interestingly, no relationship in this respect was noticed in terms of fluency. This denotes that, if all that is required of a PE task are improvements in the flow/fluency of the text, source-language knowledge is not crucial. When taking this result into account together with findings obtained in the context of Aim 1, where consulting the ST was only found to be a sign of effort for participants with a low level of proficiency in the source language, it can be concluded that not only is the fluency of the text expected to improve in a blind PE scenario (i.e. where there is no access to the ST), but also that the blind condition is expected to reduce cognitive effort if post-editors have a low level of proficiency in the source language.

6.4. Methodological Considerations on Cognitive Effort (Aim 3)

Thanks to the between-sample design adopted in this thesis it was possible to compare data obtained in the eye-tracking and think-aloud tasks with a view to estimating the degree of similarity between different measures that can be used as indicators of cognitive effort. According to results obtained in a mixed-effects statistical analysis, the count of TAP coding units was found to be positively correlated with subjective ratings on cognitive effort and with average fixation duration, measures traditionally employed for cognitive-effort estimation. These results have important implications for further research as they are testament to the fact that interferences posed by the think-aloud condition – e.g. a slow-down effect and a more fine-grained mental segmentation of the text (Jakobsen 2003; Krings 2001) – are not large enough to prevent TAPs from being used as a measure of cognitive effort. It should be noted, however, that the strength of the correlations observed here between TAPs and other measures of cognitive effort differed considerably. In particular, TAPs were more strongly associated with subjective ratings, which could be due to the fact that both subjective ratings and TAPs rely on participants' perceptions and interpretation of the task.

6.5. Cognitive Effort Revisited

As previously mentioned in this thesis, the concept of cognitive effort is notoriously elusive, both in terms of its definition and in terms of its estimation. The general argument for using this construct in PE research, as opposed to variables that can be more easily handled (e.g. PE time or number of editing operations), is based on Krings's (2001) formulation of the overall concept of effort in PE, where *cognitive* effort is regarded as the decisive variable within the triad including cognitive, technical and temporal effort (see section 2.2.1.1).

Regarding the theoretical underpinnings of cognitive effort, results obtained in this thesis gave signs that there may be more to this construct than just 'the amount of the available processing capacity' allocated to a task (Tyler et al. 1979, 608), a traditional way of defining cognitive effort in cognitive psychology. Based on the analysis carried out here, it can be postulated that cognitive effort might in fact be a multifaceted concept. Only subjective ratings and average fixation duration were found to correlate with TAPs in section 4.3, for example, with the correlation between TAPs and subjective ratings being considerably stronger. This suggests that these measures could be capturing slightly different aspects of the construct, which interestingly has also been observed in previous research regarding the concept of attention.

Attention and *effort* have on some occasions been regarded as the same concepts (Wickens 1984). As mentioned in section 2.2.2.1, attention can be defined in terms of its *control* and its *scope* (Cowan et al. 2005). Attention control concerns the capacity of selecting only a given number of items to be held in the attentional focus out of a range of possibilities, while attention scope concerns the quantity of the items kept in the attentional focus and the accuracy with which these items are held. It is interesting to note that notions of both quantity and quality (i.e. the amount and accuracy of information), a distinction based on empirical evidence (see e.g. Chow and Conway 2015), are taken into account in this formulation of attention. In an analogy with these two notions, one could think of cognitive effort perhaps not only in terms of the amount of cognitive resources expended on a task, as suggested by Tyler et al. (1979, 608), but also in terms of the intensity or quality of this expenditure.

In regard to the distinction just described, it seems that measures that are expected to vary with time alone (e.g. fixation count or total fixation time) may be closer to a notion of *amount* of cognitive effort. This is because the availability of resources in working memory is limited (Baddeley 2007; Cowan 2005; Tyler et al. 1979), which means that only so many resources can be spent at a time, resulting in demanding tasks taking longer to complete. This seems to constitute a temporal aspect of effort expenditure that may be more easily captured by time-driven measures, even when these measures are normalised (e.g. by number of characters in the ST). This potential temporal aspect of cognitive effort would also be consistent with the fact that fixation count and PE time were found to be highly correlated in the present study (see section 4.1.3.2), which denotes that the difference between temporal effort and time-driven measures used in previous research as indicators of cognitive effort may, in practice, be very hard to observe empirically.

The notion of effort *intensity* – i.e. the quality of the effort expended – may be more easily captured by subjective measures. This can be explained by the fact that information of this kind might be more easily accessed by participants than a notion of 'amount of resources', which would require recalling the 'amount of thinking' that has taken place, in quantitative terms. As subjective ratings are more holistic than other measures (O'Donnell and Eggemeier 1986), it may be that when participants provide a subjective numerical score corresponding to a given moment of the task (e.g. post-editing a specific MT sentence) this score reflects not so much the amount of thinking that went on, but a more general account of how *intensely* or how *hard* participants had to work, which is in line with the description of the mental effort dimension of the NASA-TLX scale of workload (Hart and Staveland 1988) – see section 2.2.3.2.

Even though fixation count seems closer to a notion of amount of effort and subjective ratings seem closer to a notion of effort intensity, it should be noted at this point that a clear-cut correspondence between different measures of cognitive effort and one of the two potential facets of this concept, described above, seems unlikely. Average fixation duration, for example, may be somewhere between these two notions, as this measure is normally found to differ slightly not only from other eye-movement metrics, such as fixation count (Doherty, O'Brien and Carl 2010), but also from subjective ratings (van Gog et al. 2009).

It is also worth highlighting that this potentially multifaceted nature of cognitive effort has been taken into account in this thesis through the use of a number of different data sources, which are expected to complement each other in reflecting an overall notion of cognitive effort. Results obtained in a setting of this kind provide strong reasons to suspect that studies involving the estimation of cognitive effort would benefit from a sub-division of this construct, with the notions of *intensity* and *amount* of cognitive effort, as described above, constituting two of its candidate sub-components. This would allow a more precise investigation of different cognitive phenomena whilst also shedding light on how different measures might constrain the analysis. It might be that certain cognitive tasks are predominantly associated with one of these potential facets of the construct, for example, or that certain measures constitute more suitable choices depending on the nature of the investigation. It seems that these are questions that merit further research not only in the context of PE, but in cognitive psychology in general.

6.6. Implications for Practice

In addition to the implications of specific findings of the study mentioned in the previous sections, from a more general point of view findings obtained in this thesis are also able to contribute to ongoing discussions on price estimation in PE and on the potential profile of ideal post-editors.

Regarding price estimation, findings obtained here are able to inform real-world practices in two ways: by providing information on textual predictors of cognitive effort, which can be exploited to obtain pay estimates; and by furthering the current understanding of how different indicators currently used to estimate pricing are connected to post-edited quality. In the case of the former, findings obtained in this thesis can be applied mainly within MT quality estimation – i.e. the practice of predicting MT quality based on textual characteristics of the ST and the MT output (see section 2.3.2). The different ways in which results obtained in the study can be exploited for this purpose have been explained in detail in section 5.1.1, so this is not described here again. It is worth mentioning, however, that making use of quality-estimation scores to calculate pay rates in PE has recently been described as ‘far from being a well-established technique ready for implementation’ (TAUS 2013b), which underlies the significance of research initiatives such as the one undertaken here.

Concerning the different links observed in this thesis between post-edited quality, on the one hand, and amount of changes and cognitive effort, on the other hand, it is worth signalling that cognitive effort and amount of editing are exploited in different pricing schemes currently put forth in the translation industry. A pricing scheme frequently proposed consists of estimating the degree of similarity between the raw MT output and the post-edited text and applying reductions to a pre-defined full-translation rate for sentences where few or no edits are carried out (Canek 2011). This type of approach is based entirely on editing interventions, which, as results obtained here suggest, may be unrelated to cognitive effort (see also Koponen 2012), but generally positively linked to post-edited quality (when MT quality is between low and medium). This suggests that schemes of this kind may be fairer to commissioners than to post-editors, who in this case risk not being compensated for the levels of cognitive effort invested in the task.

Asia Online (2012) propose a different approach to PE pricing. They argue that pay should be calculated according to estimated productivity gains introduced by PE relative to full translation of similar texts, suggesting, for instance, that 'if post-editing MT is 3 times faster than a human only translation […] then there is justification for reducing rates by 33% of the regular rate' (ibid.). According to the scheme proposed by Asia Online, productivity estimations should be based on the translating and post-editing behaviour of what they refer to as a 'trusted translator', i.e. 'a translator who is known to the LSP [language service provider] and who is representative of the typical translator who would be working on a project' (ibid.). Productivity levels observed for this translator would then be used to extrapolate reduction rates for the rest of the company. This approach attempts to take cognitive effort expenditure into account by paying less for less effort – assuming that effort is inversely proportional to productivity (O'Brien 2011). However, it seems to disregard the huge amounts of variation that there may be between post-editors, who risk being compensated unfairly based on this scheme because their PE operating mode might differ from that of the 'trusted translator'.

Defining the best type of effort or indicator of editing behaviour to be exploited for estimations of pay is a question that falls beyond the scope of this thesis. However, the results obtained here suggest that it is not straightforward to assume that the amount of modifications implemented in the MT output or gains in productivity obtained by a single translator will, alone, be good parameters for calculating post-editors' compensation. While modifications in the text were generally found to reflect post-edited quality, they were also found to be dissociated from cognitive effort, with extreme variations between participants being observed in this respect. This implies that the translation community may benefit from more holistic price estimation approaches. Further research could be conducted on this topic perhaps experimenting with metrics that reflect technical effort (i.e. mechanical operations) as well as the different potential facets of the cognitive effort invested in the activity, discussed above in section 6.5. Ideally, these metrics should provide estimates based on each PE task conducted, without relying on extrapolations of the behaviour of a single translator.

Regarding the potential profile of an ideal post-editor, it is worth noting that other than an expected effect of source-language knowledge on the adequacy of the post-edited text, the subjects' individual traits examined in this thesis did not seem related to post-edited quality. Indeed, recent research in this respect suggests that even individuals with no professional experience in translation or PE may be good post-editors (Mitchell, Roturier and O'Brien 2013). While results of the present study lead to a similar conclusion, a practical contribution of this thesis in this respect lies in implying that, generally, actual editing behaviour is a better index of PE expertise than post-editors' individual characteristics. This means that aspects of what post-editors actually do should be the focus of further research aimed at predicting post-edited quality and not characteristics relating to post-editors themselves.

6.7. Strengths and Limitations of the Study

To the author's knowledge, the present study constitutes the first attempt to date at combining the use of eye tracking, a self-report scale and TAPs in the context of a translation studies investigation. The mixture of these methodologies is regarded here as one of the major strengths of the study as this allows addressing a number of important issues not previously tackled in related work, such as the connections between mental processes and cognitive effort in PE and how TAPs compare to other measures that can be used as indicators of cognitive effort. This mixed-method design is also likely to be a more effective approach in estimating cognitive effort, a notoriously elusive concept which in previous research has been investigated based on isolated measures, without the contrast of methodologies adopted here.

Naturally, the comprehensive methodological approach adopted in the study also entails limitations. For example, to guarantee a reliable link between eye-tracking data and corresponding parts of the text, and to allow for a sentence-level collection of subjective ratings, participants in the eye-tracking task had to edit the MT output one sentence at a time without backtracking, a situation that deviates from a realistic professional scenario. To counterbalance this limitation, participants in the think-aloud task were exposed to the entire text at once, which made it possible to estimate the degree of interference posed by the operating mode adopted for the eye-tracking task.

TAPs were found to produce very similar findings in comparison with eye tracking (e.g. regarding correlations between the automatic evaluation metric Meteor and cognitive effort), which suggests that the restrictions imposed by the eye-tracking task did not significantly alter the nature of the results.

As the eye-tracking task was conducted on a sentence-by-sentence basis, it seemed logical to assess the post-edited texts produced in the study also at the sentence level. This too could be regarded as a limitation of the study in that there are whole-text aspects capable of influencing translation quality that could not be taken into account in the translation assessment conducted here, such as issues with document consistency and paragraph development. However, it is worth noting in this respect that post-edited quality is just one of various factors taken into account in the study, differently from cognitive effort, which is the central variable of interest in the investigation. Making use of a more comprehensive method of translation-quality assessment would have been beyond the scope of the project.

The combination of qualitative and quantitative approaches adopted in this thesis is also arguably of central relevance to the current state of research in PE and is deemed here to constitute another strength of the study. In particular, TAPs are regarded here as an extremely relevant tool capable of providing information that cannot be obtained based on other methods, such as the motivation for certain modifications implemented in the text and edits that are not physically carried out, but only mentally planned. This is a privileged angle of observation capable of revealing information that is lacking in the vast majority of previous PE research.

The use of regression analysis, carried out through mixed-effects modelling, is also regarded here as a methodological strength of the study in that a context of this kind allows for the potential impact of a number of different predictors to be concomitantly examined over and above each other's influence on the outcome variable. This avoids an incomplete analysis, i.e. investigating variations in PE effort as a function of isolated factors such as only ST features or only MT output features, a situation that is aggravated further if the potential influence of participants' individual traits is not controlled for. All of these aspects have been taken into account in the present investigation, which adds validity to the findings.

As in most empirical studies involving human subjects, results can always be influenced by the size of the participant sample. While this thesis's sample is of a relatively good size (i.e. 28 participants in total), the between-task strategy adopted here restricts the power of this sample to some extent, since, of the whole group, only nine participants took part in the think-aloud task. A small sample was considered necessary for this task in view of the laborious procedures required by the analysis of TAPs, which in the present study involved transcribing, segmenting and coding over 10 hours' worth of task time. Future research would expand this analysis, comparing eye-tracking and think-aloud data based on larger samples, with similar numbers of participants each.

Finally, it should also be mentioned that the results reported here are inevitably constrained to PE carried out from French into English, the language pair adopted for the study. Especially in regard to textual predictors of cognitive effort, findings obtained in the study should not be extrapolated to different language pairs or a different language direction, as a comparison of the results obtained here with those of previous research denotes that significant features in the context of one language may not be a factor in the context of another. In this respect, the vast majority of previous research in PE is based on English as source language, which renders the fact that English was adopted here as the *target* language a distinguishing feature of the study.

6.8. Directions for Future Research

Findings reported in this thesis provide insight into a number of aspects of PE that seem to deserve further investigation. Regarding predictions of the amounts of cognitive effort required by the activity, these aspects comprise the impact of different genres on PE effort as well as the influence that different translation aids (e.g. thesauri, term bases, search engines) may have on the levels of effort expended by participants.

As a human assessment of translation quality was only carried out in this thesis at a sentence level, further investigation could be conducted on the links between post-editing behaviour and post-edited quality, but using the text as a unit of analysis. This would allow textual aspects relating to discourse to be taken into account in the quality assessment, potentially revealing further information on how the types and amounts of effort invested in PE are connected to the levels of quality of the post-edited product.

It was also observed in the study that an intuitive editing behaviour might be more beneficial in PE than an analytical approach to the task. Here it seems that a key aspect to be tackled in future research regards the underlying mechanisms that govern different PE operating modes. From a behavioural perspective, it would be interesting to know, for example, what makes post-editors act by intuition and what makes them take a more controlled approach to the task, with characteristics of the text (ST and MT output) and of post-editors themselves being considered as potential factors in this respect.

Generally, PE seems like a very different activity when compared with traditional translation, involving what appears to be a substantially larger amount of non-linear phenomena (e.g. the fact that edits cannot be too few or too many in number). Most research to date comparing PE and fully human translation is centred on a market goal of measuring the levels of productivity and quality achieved under either condition, without providing an in-depth account of the cognitive underpinnings of these activities. Comparisons between cognitive aspects of PE and of traditional revision received even less attention, especially in empirical studies. While reviewers of human translation also need to edit an existing text, they probably have higher expectations on translation quality as well as a less critical approach to the task as they know that the text being edited has been produced by a fellow human translator. It seems then that the cognitive mechanisms underlying PE, traditional translation and traditional revision merit further investigation aimed at *understanding* these activities as opposed to examining their economic feasibility.

Appendix A

Source Texts

Sentence ID	Text A
i	Dépistage du cancer de la prostate: passer le test ou non?
ii	Depuis quelques années, des reportages dans les médias font état d’une controverse entourant le test de dépistage de l’antigène prostatique spécifique (APS), qui permet de dépister le cancer de la prostate.
iii	Récemment, un organisme gouvernemental américain, l’United States Preventive Task Force, a recommandé de ne plus faire passer systématiquement ce test aux hommes de 50 ans et plus.
1	En effet, le test d’APS présenterait parfois des résultats erronés, avec de faux résultats négatifs ou encore de faux positifs, lesquels entraînent des interventions médicales inutiles.
2	De quoi faire hésiter encore plus les hommes déjà réticents à passer des tests de dépistage.
3	Passer le test ou non?
4	Nous avons demandé l’avis de deux spécialistes.
iv	Le Dr Frédéric Pouliot, urologue, oncologue et professeur adjoint à la faculté de médecine de l’Université Laval:
5	Dans les études menées aux États-Unis, il y avait beaucoup de contamination entre les groupes témoins, il est donc difficile d’interpréter ces données et d’avoir des recommandations fermes.
6	Une autre étude, celle-là européenne, a conclu à une différence de mortalité entre les patients qui ont eu un dépistage et ceux qui n’en ont pas eu.
7	Cette étude a aussi démontré, avec un suivi après 12 ans, qu’on a entre 30 et 40% de plus de chances d’avoir des métastases si on n’est pas dépisté.
8	Je recommande donc le test à partir de 50 ans, ou à partir de 40 ans si on a un parent direct qui a déjà eu un cancer de la prostate.
9	Les hommes d’origine afro-américaine sont également plus à risque.
10	La clé est de prendre la bonne décision une fois qu’on a détecté un cancer.
11	Il y a des cancers agressifs et d’autres qui sont indolents.
12	Il faut vraiment faire comprendre au patient le degré de risque de son cancer, en lui offrant les options possibles, en ne traitant pas nécessairement les cancers de la prostate qui ne portent pas atteinte à la vie à long terme, et en optant plutôt, dans ces cas-là, pour une surveillance active de la maladie.
v	Le Dr Simon Tanguay, urologue, professeur de chirurgie à l’Université McGill et secrétaire général de la Société internationale d’urologie:
13	Aujourd’hui, beaucoup d’hommes chez qui on détecte un cancer ne

seront pas traités, car leur cancer n’est pas agressif et ne menace pas leur vie.

14 On va leur suggérer de faire de la surveillance active et si la maladie progresse, on va leur offrir un traitement.

15 De plus en plus, on détermine avec précision des critères pour décider qui devrait ou ne devrait pas être traité.

16 Je recommande donc quand même de passer le test.

17 Mais l’important est d’avoir une discussion avec son médecin pour déterminer si on devrait le passer ou non.

18 En collaboration avec la Société internationale d’urologie, Movember a créé un outil qui permet d’évaluer le pour et le contre du test d’APS.

19 On peut télécharger ce document (en anglais pour l’instant, une traduction sera offerte sous peu) à cette adresse: http://ca.movember.com/fr/mens-health/prostate-cancer-screening

Text B

vi Une stratégie républicaine pour contrer la réélection d’Obama

vii Depuis janvier 2009, les leaders républicains au Congrès ont défini une stratégie pour s’assurer que la présidence de Barack Obama se limite à un mandat.

viii Un des éléments clefs de cette stratégie consiste à parrainer dans plus de 40 États des lois électorales visant à restreindre le droit de vote.

ix L’objectif est clair : empêcher une victoire de Barack Obama en 2012 en limitant la participation électorale.

20 Les dirigeants républicains justifièrent leur politique par la nécessité de lutter contre la fraude électorale.

21 Or, le Centre Brennan considère cette dernière comme un mythe, affirmant que la fraude électorale est plus rare aux États-Unis que le nombre de personnes tuées par la foudre.

22 D’ailleurs, les avocats républicains n’ont recensé que 300 cas de fraude électorale aux États-Unis en dix ans.

23 Une chose est certaine: ces nouvelles dispositions influenceront négativement le taux de participation.

24 En ce sens, ces mesures mineront en partie le système démocratique américain.

25 Contrairement au Canada, les États américains sont responsables de l’organisation des élections fédérales aux États-Unis.

26 C’est dans cet esprit qu’une majorité de gouvernements américains promulguèrent à partir de 2009 de nouvelles lois rendant plus difficile le processus d’inscription ou de vote.

27 Ce phénomène a pris de l’ampleur après les élections de novembre 2010 qui virent s’ajouter 675 nouveaux représentants républicains dans 26 États.

28 En conséquence, 180 projets de lois restreignant l’exercice du droit de vote dans 41 États furent introduits durant la seule année de 2011.

29 Les nouvelles lois électorales exigent que les électeurs présentent une carte d’identité avec photo et une preuve de citoyenneté américaine.

30 Par ailleurs, ces lois réduisent aussi les périodes de vote par anticipation, invalident le droit de s’inscrire comme électeur le jour du scrutin et retirent aux citoyens ayant un dossier judiciaire leur droit de vote.

31 Avant les élections de 2006, aucun État américain n’exigeait des électeurs de présenter une carte d’identité avec photo.

32 L’Indiana fut le premier État à poser une telle exigence.

33 La Cour Suprême des États-Unis confirma en 2008 la constitutionnalité de la loi de l’Indiana.

34 Les autorités républicaines s’empressèrent d’étendre cette pratique à d’autres États.

35 Au cours des deux dernières années, elles parrainaient des projets de loi dans 34 États pour forcer les électeurs à présenter une carte d’identité avec photo.

36 Il est important de noter que, contrairement au Québec, les citoyens américains ne disposent pas de carte d’identité universelle comme la carte de l’assurance maladie.

37 De fait, 11% des citoyens américains, soit 21 millions de personnes en âge de voter, ne possèdent pas de cartes d’identité avec photo émises par une agence gouvernementale de leur État.

38 Par ailleurs, cinq millions de nouveaux électeurs en 2012 ne disposent pas d’une telle pièce d’identité.

39 Or, il en coûte souvent plus de cent dollars pour obtenir la carte d’identité requise.

40 Les nouvelles restrictions affectent de manière disproportionnée les jeunes, les minorités, et les personnes à faible revenu.

41 En effet, 25% des Afro-Américains, 15% des personnes gagnant moins de 35 000 dollars; 18% des citoyens de plus de 65 ans et 20% des électeurs de 18 à 29 ans ne possèdent pas la carte d’identité requise avec photo.

Machine Translations

Sentence ID **Text A**

i Screening of the cancer of the prostate: to pass the test or no?

ii In recent years, media reports indicate a controversy surrounding the screening test for prostate specific antigen (PSA), which can detect cancer of the prostate.

iii Rcemment, an organism governmental amricain, the States United Preventive Task Force, has recommand of more to make pass systmatiquement this test to the men of 50 years and more.

1 Indeed, the test of APS prsenterait sometimes of the rsultats wander, with false rsultats ngatifs or false positives, which entranent of interventions useless mdicales.

2 Of what to do hsiter again more the men dj rticents to pass tests of dpistage.

3 Take the test or not?

4 We have asked the opinion of two specialists.

iv The Dr Frdric Pouliot, urologist, oncologist and professor attaches the facult of mdecine of the Laval Universit:

5 In studies conducted in the United States, there was a lot of contamination between groups, so it is difficult to interpret the data and have firm recommendations.

6 Another study, this one european, concluded that there was a difference in mortality between patients who have had a screening and those who have not.

7 The study also showed, with a follow-up after 12 years, that it was between 30 and 40% more likely to have metastasis if it is not detected.

8 I recommend the test therefore to leave from 50 years, or to leave from 40 years if one has a direct parent who has dj had a cancer of the prostate.

9 The African-American men are also more at risk.

10 The key is to make the right decision once cancer has been detected.

11 There are aggressive cancers and others which are indolents.

12 We really need to understand the patient the degree of risk of the cancer, by offering options, by not addressing necessarily prostate cancer who do not affect life in the long term, and opting instead, in these cases ,-for active surveillance of the disease.

v Dr. Simon Tanguay, urologist, professor of surgery at McGill University and secretary general of the International Society of urology:

13 Today, many men in whom cancer is detected will not be treated because their cancer is not aggressive and does not threaten their lives.

14 It will suggest to active monitoring and if the disease progresses, it will provide a treatment.

15 More and more, one dtermine with prcision of the critres for dcider that would have to or would not owe tre milks.

16 I would recommend it to the test.

17 But the important is to have a debate with the doctor to determine if they should pass him or not.

18 In collaboration with the international society of Urology, Movember has created a tool that allows to evaluate the pros and cons of the PSA test.

19 You can download this document (in English for the moment, a translation will be provided shortly) to this address: http://ca.movember.com/fr/mens-health/prostate-cancer-screening

Text B

vi A republican strategy to oppose the reelection of Obama

vii Since January 2009, the Republican leaders in Congress have defined a strategy to ensure that the Presidency of Barack Obama is limited to a mandate.

viii One of the elements keys of this strategy consist in to sponsor in more than 40 States of the electoral laws aiming to restrict the vote right.

ix The objective is clear: to prevent a victory of Barack Obama in 2012 while limiting electoral participation.

20 The Republican leaders justified their policies by the need to combat electoral fraud.

21 However, the Brennan Center considers the latter as a myth, stating that electoral fraud is rarer in the United States than the number of people killed by lightning.

22 Indeed, Republican lawyers have identified only 300 cases of electoral fraud in the United States in a decade.

23 A thing is certain: this new arrangements will influence ngativement the rate of involvement.

24 In this sense, these measures will mine the systme in part dmocratique amricain.

25 Unlike Canada, the American States are responsible for the organization of federal elections in the United States.

26 It is in this spirit that a majority of governments americans promulguèrent from 2009 new laws making it more difficult the process of registration or voting.

27 This phnomne took of the size aprs the lections of November 2010 that turns to be added 675 new reprsentants rpublicains in 26 tats.

28 As a result, 180 draft laws restricting the exercise of the right to vote in 41 states were introduced during the single year of 2011.

29 The new Elections Acts demand that the voters introduce an identity card with photo and a proof of American citizenship.

30 In addition, these laws also reduce the periods of early voting, invalidate the right to register as a voter on election day and withdraw to citizens with a criminal record their right to vote.

31 Before the elections of 2006, no U.S. state did voters to present a photo identification card.

32 Indiana was the first state to impose such a requirement.

33 In 2008, the Supreme Court of the United States confirmed the constitutionality of the Indiana law.

34 The autorits rpublicaines himself empressrent to stretch this practice of other tats.

35 Over the past two years, sponsors of the bill in 34 states to force voters to present an identity card with a photograph.

36 It is important to note that, unlike Quebec, American citizens are not universal ID such as the health insurance card.

37 In fact, 11% of the citizens amricains, either 21 millions of people in ge to vote, possdent not of cards of identit with photo stakes by a governmental agency of their tat.

38 In addition, five million new voters in 2012 do not have such identification.

39 However, it often costs more than $100 for the ID card required.

40 The new restrictions allocate the youths, the minorities, and the people in a disproportionate way to weak income.

41 In fact, 25% of African-americans , 15% of the people earning less than 35,000 dollars; 18% of the citizens of more than 65 years and 20% of the electorate in 18 to 29 years do not have the identity card required with photo.

KEY:

i-ix: Not considered for quantitative analysis

1-41: Considered for quantitative analysis

Appendix B

Post-Editing Brief

Many thanks for agreeing to take part in this study!

You will edit English translations of two news articles originally published in French. You will be able to see the French original text whilst editing the translation, which was produced automatically with Machine Translation software.

Before carrying out the task, please consider the instructions below:

1. You should try and render the text resulting from your editing work acceptable for publication in an English-speaking context, i.e. the translation should be accurate, the text should be grammatically correct, stylistically fine, and read well to an English native speaker.

2. Professional guidelines advise that you should keep as much of the existing translation as possible, i.e. unnecessary changes are to be avoided. If the text is correct and you think it sounds fine, you should not substitute words for alternatives that you 'like' better.

3. Real-world post-editing involves processing large amounts of text under fairly tight deadlines. Hence, bearing in mind instruction 1, you should try and finish editing each segment as fast as you can.

4. If you don't know the meaning of a word, feel free to make informed guesses. However, when the French does not help and you are absolutely clueless, you should trust the existing translation.

If you have further questions, please don't hesitate to ask me! ☺

Appendix C

Translation Quality Assessment Brief

Many thanks for agreeing to take part in this project!
You will evaluate translations of two French news articles. Before starting the evaluation, please read the two short texts below, which may aid in the assessment.

Background information for article A:

Prostate cancer is a form of cancer that develops in the prostate, a gland in the male reproductive system.

PSA (prostate-specific antigen) is present in small quantities in the serum of men with healthy prostates, but is often elevated in the presence of prostate cancer or other prostate disorders. The United States Preventive Services Task Force (USPSTF) does not recommend PSA screening, noting that the test may result in “overdiagnosis” and “overtreatment” because “most prostate cancer is asymptomatic for life,” and treatments involve risks of complications.

Clinical practice guidelines for prostate cancer screening vary and are controversial due to uncertainty as to whether the benefits of screening ultimately outweigh the risks of overdiagnosis and overtreatment.

Movember is an annual, month-long event involving the growing of moustaches during the month of November to raise awareness of prostate cancer and other male cancer and associated charities. The Movember Foundation runs the Movember charity event, housed at Movember.com. The goal of Movember is to “change the face of men’s health.”

With thanks to Wikipedia©

Background information for article B:

The Republican Party is one of the two major contemporary political parties in the United States, the other being the Democratic Party.

The Brennan Center is a public policy and law institute that focuses on issues involving democracy and justice. It is involved in issues such as voting rights, redistricting reform, campaign finance reform, and presidential power in the fight against terrorism.

The United States has a federal government, with elected officials at the federal (national), state and local levels. A state of the United States of America is one of the 50 constituent political entities that shares its sovereignty with the United States federal government. State law regulates most aspects of the election, including primaries, the eligibility of voters (beyond the basic constitutional definition), the running of each state’s electoral college, and the running of state and local elections.

There is no true national identity card in the United States of America, in the sense that there is no federal agency with nationwide jurisdiction that directly issues such cards to all American citizens for mandatory regular use. All legislative attempts to create one have failed due to tenacious opposition from liberal and conservative politicians alike, who usually regard the national identity card as the mark of a totalitarian society.

With thanks to Wikipedia©

If possible, also take the time to read the original articles online by following the links provided below.

Article A:

http://www.lapresse.ca/vivre/sante/201211/30/01-4599309-depistage-du-cancer-de-la-prostate-passer-le-test-ou-non.php

Article B:

http://www.lapresse.ca/la-tribune/opinions/201207/30/01-4560667-une-strategie-republicaine-pour-contrer-la-reelection-dobama.php

Please, follow the link below to access reference translations of these articles previously carried out by professional translators. Note that the purpose of this reference is to provide you with an idea of what reasonably good translations of these articles should sound like. You should NOT refer to the reference once you start rating the sentences. http://forms.ncl.ac.uk/view.php?id=6058

The sentences should be rated in terms of both **Adequacy** and **Fluency**.

Adequacy captures to what extent the meaning in the source text is also expressed in the translation.

Rating scale:

4. Everything All the meaning in the source is contained in the translation, no more, no less.
3. Most Almost all the meaning in the source is contained in the translation.
2. Little Fragments of the meaning in the source are contained in the translation.
1. None None of the meaning in the source is contained in the translation.

Fluency captures to what extent the translation is well formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker.

Rating scale:

4. Flawless Refers to a perfectly flowing text with no errors.
3. Good Refers to a smoothly flowing text even when a number of minor errors are present.
2. Disfluent Refers to a text that is poorly written and difficult to understand.
1. Incomprehensible Refers to a very poorly written text that is impossible to understand.

Feel free to leave any comments in the space provided if you so wish.

Scoring examples:

Au cours des dernières années, plusieurs scientifiques ont étudié les liens entre les suppléments vitaminiques et le cancer.

In recent years, a number of scientists have studied the relationship between vitamin supplements and cancer.

Adequacy: 4 / Fluency: 4

OBS: The translation is fluent and conveys all the meaning in the source text.

Ces restrictions ne sont pas sans conséquence.

These restrictions do not have any consequence.

Adequacy: 1 / Fluency: 4

OBS: Albeit fluent, the translation conveys the opposite meaning of the original.

L'effet de la vitamine D sur le cancer n'est pas clairement établi non plus.

The effect of vitamin D on the cancer is not clear either.

Adequacy: 4 / Fluency: 3

OBS: All meaning in the source is also conveyed in the translation, but the use of the definite article ('the') with 'cancer' would constitute a minor fluency error.

Plusieurs projets de loi ont été bloqués par les vetos des gouverneurs démocrates.

Several projects of law have t bloqus by the vetoes of the governors dmocrates.

Adequacy: 2 / Fluency: 1

OBS: The translation as a whole is incomprehensible, hence the fluency score of 1. Fragments of the original meaning are, however, present – e.g. it is possible tell that something about 'several projects of law' is being said – hence the adequacy score of 2.

Additional observations

- Inverted commas were removed from all sentences due to methodological constraints. You should not take the absence of inverted commas into account in your assessment.

- If translations for the French sentence below do not include the content between brackets, this should not be taken into account in your assessment either.

On peut télécharger ce document (en anglais pour l'instant, une traduction sera offerte sous peu) à cette adresse: http://ca.movember.com/fr/mens-health/prostate-cancer-screening

Please note that you should try and do each batch in one sitting. You can do Internet searches and consult other materials if you feel the need to. There is no limit for breaks taken between batches, as long as two batches are completed each day, on consecutive days. You should do the batches in the order you receive them. You will be sent separate emails, each containing a link to its respective batch – i.e. six emails in total. The interface you will be using looks like this:

You will always be assessing the sentences in the green background – also being able to see the preceding (if any) and the following sentence. You should try and rate the translations' fluency first, without looking at the French. After choosing a level for fluency, you can compare the translation with the original and choose an adequacy level and then just click next to go to the next sentence.

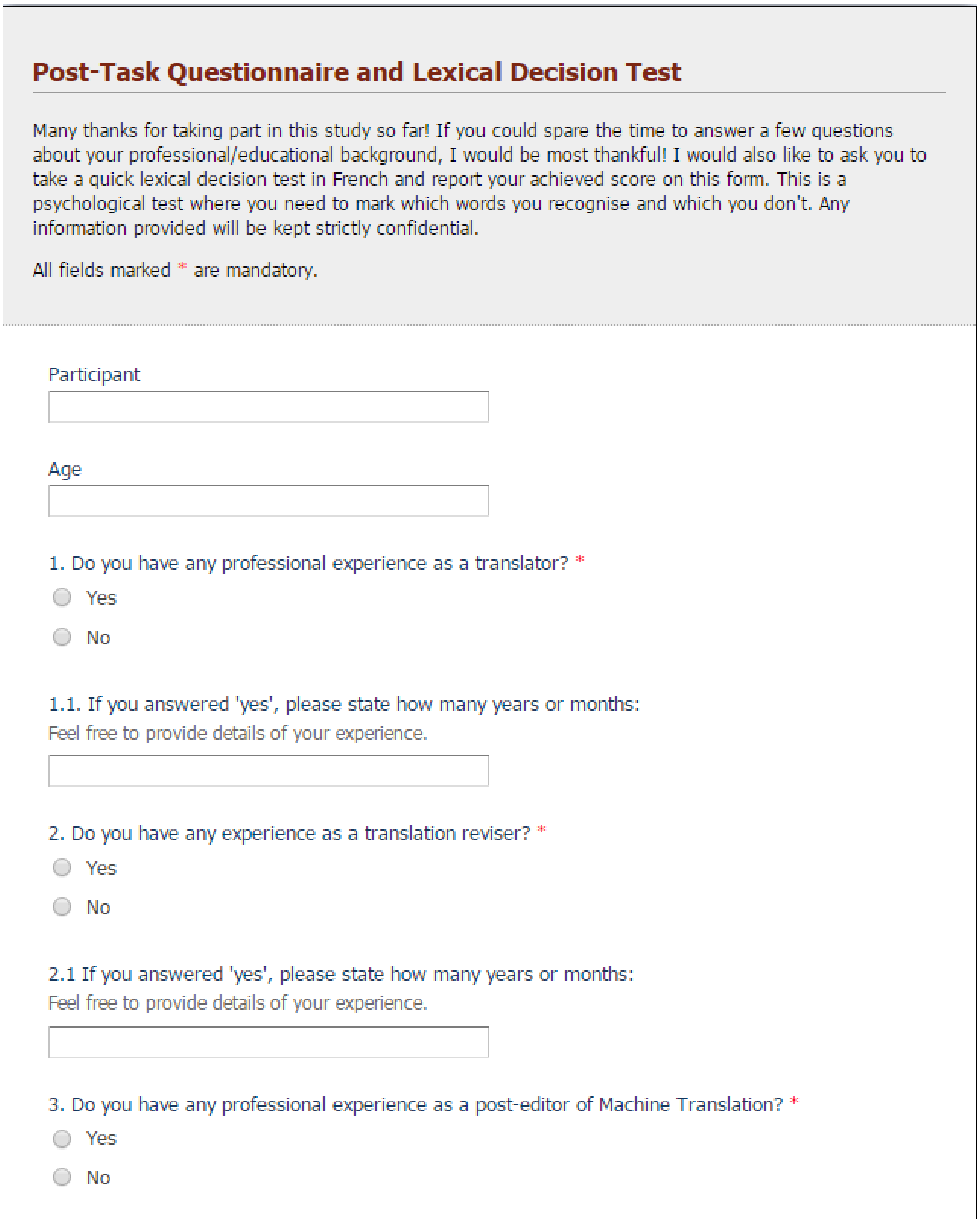
Feel free to email me if you have any questions.

Once again, many thanks!

Lucas N Vieira

Appendix D

Post-Task Questionnaire



Post-Task Questionnaire and Lexical Decision Test

Many thanks for taking part in this study so far! If you could spare the time to answer a few questions about your professional/educational background, I would be most thankful! I would also like to ask you to take a quick lexical decision test in French and report your achieved score on this form. This is a psychological test where you need to mark which words you recognise and which you don't. Any information provided will be kept strictly confidential.

All fields marked * are mandatory.

Participant

Age

1. Do you have any professional experience as a translator? *

○ Yes

○ No

1.1. If you answered 'yes', please state how many years or months:

Feel free to provide details of your experience.

2. Do you have any experience as a translation reviser? *

○ Yes

○ No

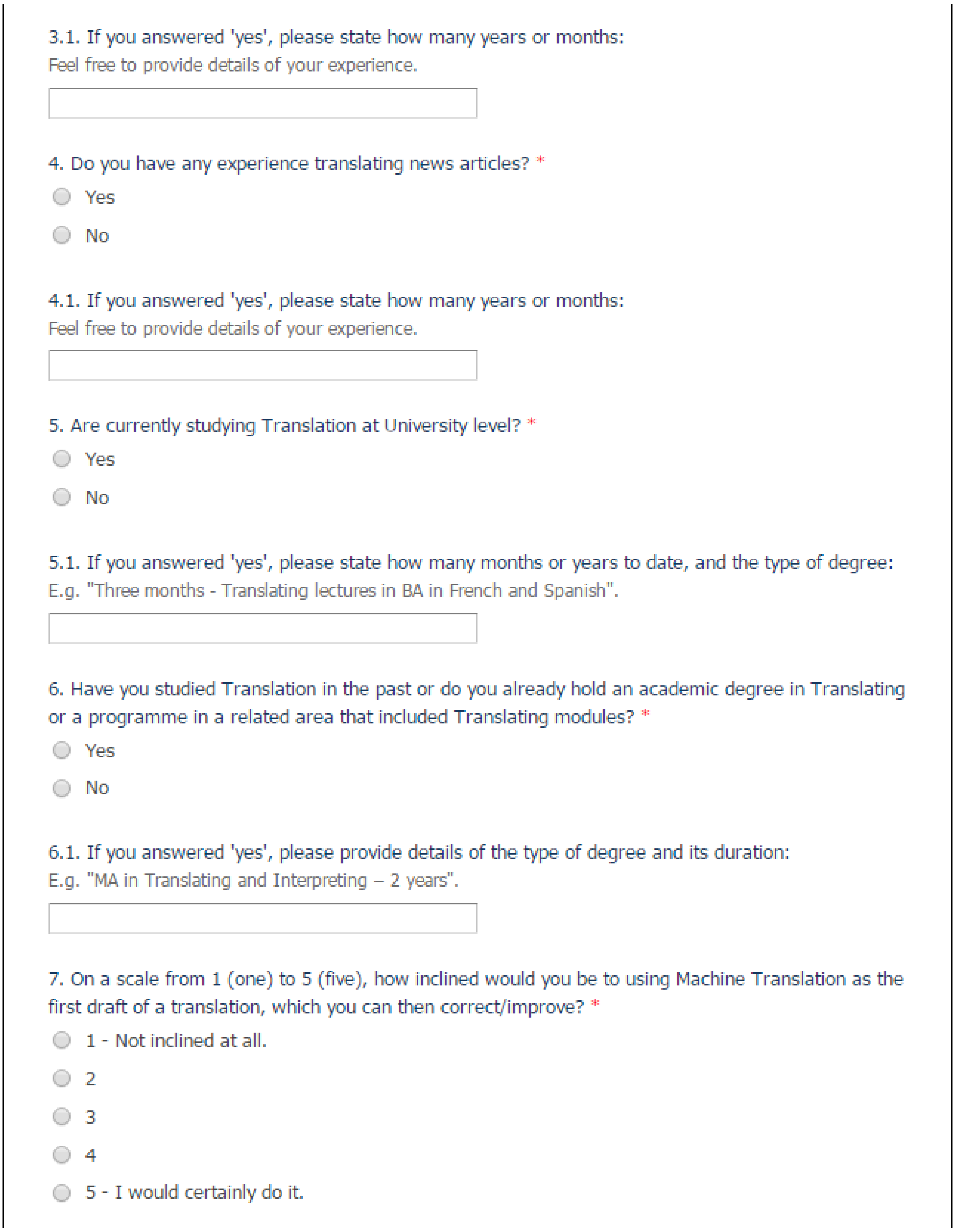
2.1 If you answered 'yes', please state how many years or months:

Feel free to provide details of your experience.

3. Do you have any professional experience as a post-editor of Machine Translation? *

○ Yes

○ No



3.1. If you answered 'yes', please state how many years or months:

Feel free to provide details of your experience.

4. Do you have any experience translating news articles? *

○ Yes

○ No

4.1. If you answered 'yes', please state how many years or months:

Feel free to provide details of your experience.

5. Are currently studying Translation at University level? *

○ Yes

○ No

5.1. If you answered 'yes', please state how many months or years to date, and the type of degree:

E.g. "Three months - Translating lectures in BA in French and Spanish".

6. Have you studied Translation in the past or do you already hold an academic degree in Translating or a programme in a related area that included Translating modules? *

○ Yes

○ No

6.1. If you answered 'yes', please provide details of the type of degree and its duration:

E.g. "MA in Translating and Interpreting – 2 years".

7. On a scale from 1 (one) to 5 (five), how inclined would you be to using Machine Translation as the first draft of a translation, which you can then correct/improve? *

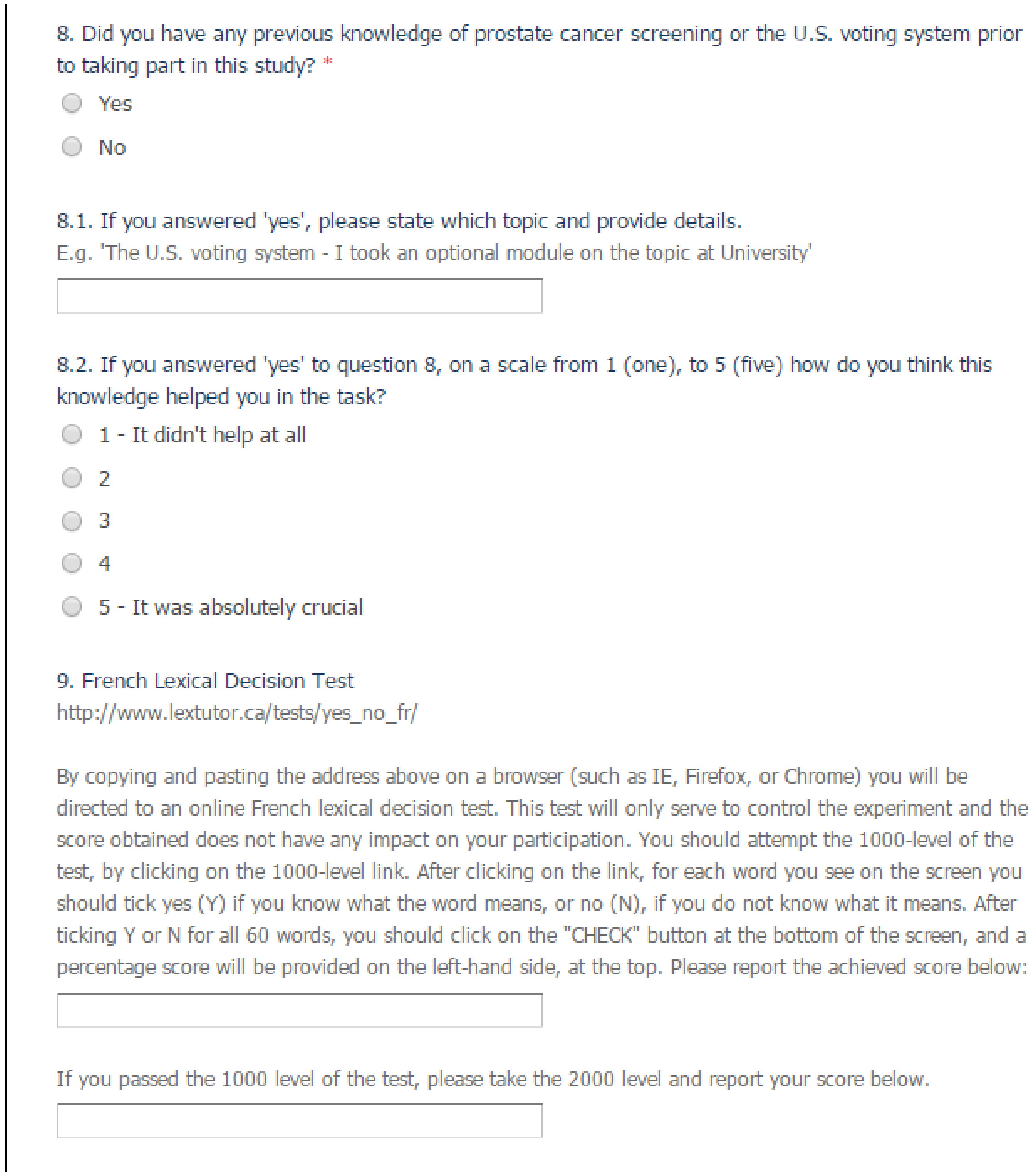
○ 1 - Not inclined at all.

○ 2

○ 3

○ 4

○ 5 - I would certainly do it.



8. Did you have any previous knowledge of prostate cancer screening or the U.S. voting system prior to taking part in this study? *

○ Yes

○ No

8.1. If you answered 'yes', please state which topic and provide details.

E.g. 'The U.S. voting system - I took an optional module on the topic at University'

8.2. If you answered 'yes' to question 8, on a scale from 1 (one), to 5 (five) how do you think this knowledge helped you in the task?

○ 1 - It didn't help at all

○ 2

○ 3

○ 4

○ 5 - It was absolutely crucial

9. French Lexical Decision Test

http://www.lextutor.ca/tests/yes_no_fr/

By copying and pasting the address above on a browser (such as IE, Firefox, or Chrome) you will be directed to an online French lexical decision test. This test will only serve to control the experiment and the score obtained does not have any impact on your participation. You should attempt the 1000-level of the test, by clicking on the 1000-level link. After clicking on the link, for each word you see on the screen you should tick yes (Y) if you know what the word means, or no (N), if you do not know what it means. After ticking Y or N for all 60 words, you should click on the "CHECK" button at the bottom of the screen, and a percentage score will be provided on the left-hand side, at the top. Please report the achieved score below:

If you passed the 1000 level of the test, please take the 2000 level and report your score below.

Appendix E

Participant Information Sheet

I am carrying out a project exploring how human editors/translators edit Machine Translation output. In order to do so, I need to recruit participants to edit/improve translations produced with computer software.

What does the task involve, exactly?

I will meet you in the University at an agreed date and time and ask you to edit a few sentences. The computer used will be connected to an eye tracker, which will record your eye movements – where you look on screen and for how long. This will only affect the way you use the computer with regard to the distance you will be sitting from the screen – no wires or equipment of any sort will be attached to your body. The software used will keep logs of your keyboard and mouse activity, and I might also ask you to describe the editing process. The entire task should take one hour and a half, on average, and it comprises the editing of two short news articles and an experiment where you will have your memory tested. You will be able to take a ten-minute break halfway through the task, when tea and coffee will be provided. After editing the texts I will also ask you to fill out a post-task questionnaire, which takes an average of 8 minutes to be completed, in total, and includes a quick lexical decision task. Any data produced will be used for research purposes only and will be kept strictly confidential.

Do I have to participate?

No. **Participation in the project is strictly voluntary**.

Do I gain anything if I choose to participate?

Payment is not offered but, as a way of saying thanks, you will be given a £10 book voucher for your participation.

What will be done with my data?

Any data produced as a result of your participation will be kept confidential and no mention will be made to your real identity. The data will be securely stored and, prior to being anonymized, will only be accessible to me.

Are there any risks involved?

The eye tracker used in the experiment is a safe device largely used in research and in the advertising industry and it will cause no harm to your eyes or sight. No significant risks will be posed as a result of computer use either, since you will be using the computer for no longer than an hour and a half.
You will also be told the safety procedures to follow in case of an emergency.

What do I do now?

Consider the information on this sheet and write to me if you have any questions. If you have no questions, I will see you in the eye-tracking lab!

HOPEFULLY THANK YOU VERY MUCH FOR YOUR HELP!

Lucas Nunes Vieira

Appendix F

Consent Form

Many thanks for considering taking part in this project!

NOTE: This consent form will remain with the researcher for his records. A copy will also be left with you.

I understand I have been asked to take part in the research project specified in the information sheet that has been presented to me.

I understand that:	**YES**	**NO**
- I shall be asked to edit texts produced by a Machine Translation system whilst having my eye movements recorded by an eye tracker	□	□
- I shall also have my keyboard and mouse activity recorded	□	□
- I shall be asked to complete questionnaires with questions about the editing task and/or about my educational/professional background	□	□
- I may be asked to comment on the experiment after and/or during the task	□	□
- I shall be asked to take language vocabulary and Working Memory tests	□	□
- unless I otherwise inform the researcher before the task, I agree to allow the screen to be videoed and my voice to be audio-recorded	□	□

and

I understand that my participation is voluntary and that I can withdraw at any stage of the project without being penalised or disadvantaged in any way;

and

I agree that, if I am a student at Newcastle University, the only compensation I shall receive for taking part in the project is training in Machine Translation Post-Editing and a £10 book voucher, and that I will not be advantaged in any way over other students that are not taking part in the research;

and

I understand that I may ask for my data to be withdrawn from the project at any time prior to publication;

and

I understand that data from eye tracking, keyboard/mouse logging, audio recording, transcripts, and questionnaires will be kept in secure storage and will be accessible only to the researcher and research supervisors;

and

I understand that data from eye tracking, keyboard/mouse logging, audio recording, transcripts and questionnaires will be kept confidential and that no reference will be made to my real identity in any reports, publications or presentations.

Please complete the following:	**YES**	**NO**	**N/A**
- Are you knowingly susceptible to epileptic seizures when exposed to light, such as the light normally emitted by the computer screen?	□	□	
- Have you read the information on this form?	□	□	
- Do you understand the information provided?	□	□	
- Have you had an opportunity to ask questions and discuss this study?	□	□	
- Have you received satisfactory answers to all your questions, if you had any?	□	□	□
- Do you agree to be contacted after taking part in the experiment should the researcher need clarifications on the data produced?	□	□	
- Do you agree to take part in further experiments in future phases of the research?	□	□	

Participant’s name in BLOCK CAPITALS: ________________________

Participant’s signature: ________________________

Witness: ________________________

Date: ________________________

Appendix G

Meteor scores and per-sentence raw counts of POS features used for the analysis presented in section 4.1.2

sentence_ID	Meteor	N	V	ADJ	ADV	P	VP	NP	AP	PP
1	0.207203	6	2	7	2	4	0	6	5	2
2	0.156258	2	3	1	3	2	3	3	1	1
3	1	1	1	0	1	0	1	1	0	0
4	0.375712	2	2	0	0	1	0	2	0	1
5	0.617654	7	5	3	2	6	3	6	2	3
6	0.523688	4	6	2	3	3	0	9	1	3
7	0.427559	7	6	0	4	6	1	7	0	4
8	0.352602	6	6	1	2	5	0	7	1	3
9	0.607999	3	1	2	2	2	0	3	1	2
10	1	4	4	1	0	1	1	4	0	0
11	0.888889	1	2	2	0	0	0	3	2	1
12	0.389963	12	5	3	8	10	3	10	2	6
13	0.540737	5	5	1	7	2	0	6	1	1
14	0.414135	3	6	1	0	2	3	3	0	1
15	0.140357	2	6	0	4	4	2	3	0	1
16	0.283467	1	2	0	3	1	1	1	0	0
17	0.307906	2	5	1	1	3	3	3	0	1
18	0.422116	7	4	1	1	7	1	9	0	4
19	0.508845	6	4	0	1	4	1	5	0	3
20	0.611102	4	2	2	0	3	1	4	1	2
21	0.621387	7	3	4	1	4	2	8	2	4
22	0.615375	7	2	2	2	4	0	5	1	3
23	0.270905	4	1	3	1	1	0	4	2	1
24	0.239144	4	1	2	0	2	0	2	2	0
25	0.526179	6	1	3	1	4	0	6	2	4
26	0.394915	8	3	3	1	7	1	8	2	6
27	0.194961	8	4	2	0	4	1	7	0	4
28	0.523912	9	3	1	0	7	1	7	0	5
29	0.360676	7	2	3	0	3	0	7	2	3
30	0.511717	13	4	1	1	8	2	11	1	5
31	0.41418	7	2	1	1	5	1	7	1	4
32	1	3	2	2	0	1	1	3	0	0
33	0.955556	6	1	2	0	4	0	6	0	4
34	0.190543	3	2	2	0	2	1	3	1	1
35	0.431812	9	3	1	0	8	2	8	0	5
36	0.578812	7	3	3	3	6	1	6	2	5
37	0.263276	11	3	2	2	10	2	10	2	8
38	1	6	1	2	2	5	0	5	0	4
39	0.283847	3	2	1	2	3	1	3	1	1
40	0.260343	6	1	3	0	2	0	5	1	1
41	0.463908	16	2	2	4	12	1	14	1	9

Principal components used for the analysis presented in section 4.1.2

sentence_ID	PC1	PC2	PC3	PC4	PC5	PC6	PC7
1	-0.32498	2.73498	-0.75629	1.159768	-0.19239	-0.4795	-0.11367
2	1.259024	2.982526	-0.1528	-1.16367	-0.75022	-0.6228	1.088565
3	-1.00799	0.846818	1.208795	-2.93369	-1.48587	-0.49375	-0.44282
4	0.81578	-3.39437	2.494873	0.327853	-2.41364	-1.12099	0.409889
5	-0.65484	-1.79309	-0.169	-0.01964	0.710484	0.515283	0.058099
6	-2.89265	-1.55697	-0.50033	0.22302	0.318919	-0.00142	-0.56851
7	-3.48749	-1.07024	-0.97015	0.906914	0.203343	0.37243	-0.24269
8	-1.15984	-0.38841	-1.40572	0.300735	0.072385	-0.16855	-1.03157
9	-1.35671	0.059735	1.435638	-1.20445	1.556476	-1.18194	1.713012
10	-0.62476	-0.53458	-1.29425	-4.23234	-0.77044	-0.12465	-0.9515
11	-1.94051	0.984193	0.122656	0.033671	-1.98064	1.790533	-0.19266
12	1.181466	-0.21719	-1.15317	0.00625	-0.06888	-0.15271	0.426454
13	-0.22832	1.227699	-1.25085	0.099533	-0.95595	-0.23983	-1.01596
14	1.613162	-1.95086	0.002307	1.563244	-2.00738	-3.08101	-0.12322
15	-2.84744	0.882136	-0.4703	0.860346	-0.63169	-0.0432	1.529782
16	0.326229	1.057514	0.580944	0.65435	-1.31946	1.656101	0.686527
17	-1.87847	-2.2681	-0.36547	-0.39387	-0.16289	1.34798	0.96096
18	1.967935	-0.90709	-0.46171	0.445877	-0.2187	1.15871	-0.06063
19	0.093585	-0.81314	2.648535	0.383342	1.163105	0.330931	0.607299
20	1.941959	0.784389	2.230621	1.170823	0.971952	0.568712	-1.48984
21	0.699028	-0.36552	0.211497	0.163856	0.289613	0.843555	0.143357
22	2.48232	0.47679	1.618695	-0.58649	-0.32316	0.679617	-0.33072
23	-1.00315	-0.01621	1.931676	-0.94031	1.673791	-0.09494	-0.46474
24	0.183479	-0.28442	-0.12778	-0.49393	1.179538	-0.51957	0.763924
25	-0.06532	0.267685	0.840084	-0.23643	0.505074	0.070567	1.329423
26	-0.65993	-0.98133	-0.67099	-0.84746	1.171021	-0.35216	0.144358
27	-1.34263	0.882862	-0.80452	-0.32505	-0.30848	0.088402	-0.61116
28	0.31093	-0.41916	-1.84651	-0.00841	0.822471	-0.48078	-0.46816
29	3.078568	0.078365	-0.0566	-0.30827	0.7886	-0.21924	-0.97046
30	1.536634	-0.34486	-0.58303	-0.22105	0.58137	-0.40764	-0.08886
31	2.403295	-1.04663	0.107981	-0.16928	0.118787	0.901614	-0.81037
32	-0.97234	1.699634	0.914346	0.204886	-0.10859	-0.20941	-0.51216
33	1.512061	0.036707	-1.6782	0.614632	-0.43283	0.677081	1.174112
34	-3.06881	1.980801	1.979357	1.46383	0.336706	-1.06947	-1.03931
35	2.928588	0.724402	-0.24721	-0.47649	-0.36752	0.36037	0.634168
36	-0.48399	-0.33676	-0.18241	0.594997	0.367659	0.054246	-0.50508
37	2.275954	-0.05858	-0.3175	1.293423	0.106513	0.541874	0.022505
38	-1.27487	-1.89785	0.100448	0.509786	0.150619	0.584465	-0.48885
39	-0.94548	1.667256	0.828771	0.803295	-0.40562	0.053832	0.261059
40	1.665445	1.063381	-0.93208	-0.30179	1.323056	-0.77761	0.352527
41	-0.05493	0.207475	-2.86035	1.078183	0.492867	-0.75513	0.216901

Sample of variables used for the analyses presented in Table 10 and Table 12 – values correspond to P01 (eye-tracking task) and J01

sentence		fix	subj.	J01	J01	J01	J01
ID	hter	(norm.)	cogeffort	FluencyMT	FluencyPE	AdequacyMT	AdequacyPE
1	0.61538	1.308108	5	1	3	1	4
2	1	3.793478	8	1	1	1	2
3	0.14286	0.545455	3	4	4	4	4
4	0	0.404255	2	3	3	4	4
5	0.0303	0.407407	2	4	4	3	4
6	0	0.398649	3	3	3	4	4
7	0.30303	1.912162	6	3	3	2	4
8	0.44444	1.541353	5	2	3	2	4
9	0.1	0.484848	3	3	4	4	4
10	0	0.333333	2	4	4	4	4
11	0.15385	0.516667	3	3	4	3	3
12	0.19231	0.980645	6	2	3	3	4
13	0	0.248276	2	4	4	4	4
14	0.52174	0.852174	4	2	3	2	4
15	0.80952	1.60177	5	1	3	2	3
16	0.3	0.5625	3	1	4	1	3
17	0.26087	0.712963	5	2	3	1	2
18	0.11111	0.390071	4	3	3	4	4
19	0.02632	0.252809	2	3	4	3	4
20	0.07143	0.504505	3	4	4	4	4
21	0	0.396552	2	3	3	4	4
22	0	0.336364	2	4	4	4	4
23	0.46667	0.864078	5	2	3	3	3
24	0.42857	1.558442	5	1	2	2	3
25	0.05556	0.545455	3	3	3	4	4
26	0.32	1.4375	6	1	1	2	2
27	0.52381	1.524138	7	1	3	2	3
28	0.17391	0.654676	4	3	3	4	3
29	0.09524	0.671429	4	3	3	3	3
30	0.18919	0.793578	6	3	3	4	3
31	0.41176	1.158333	5	2	4	3	1
32	0	0.315789	2	4	4	4	4
33	0	0.494624	2	4	4	4	4
34	0.72727	0.835294	4	1	3	1	2
35	0.12	0.64557	4	2	4	2	3
36	0.17391	0.592814	5	2	3	1	4
37	0.42857	1.086486	6	1	2	2	3
38	0	0.240385	3	4	4	4	4
39	0	0.453488	2	4	4	3	3
40	0.5	0.856	4	2	4	1	4
41	0.21739	0.785047	4	2	3	3	4

Sample of variables used for the analyses presented in sections 4.2 and 4.3 – values correspond to P20 (think-aloud task)

sentenceID	TAP units	words verbalised	TAP_sequences (micro)	Grammar units	Lexis units
1	24	69	8	4	18
2	16	73	3	3	10
3	2	8	1	1	0
4	2	16	1	0	0
5	12	84	3	0	8
6	6	30	2	4	1
7	13	90	3	13	0
8	17	73	4	4	10
9	4	18	2	1	0
10	1	13	1	0	0
11	10	47	2	1	6
12	39	175	5	14	15
13	8	37	1	2	0
14	10	56	2	5	4
15	16	84	3	6	4
16	9	36	1	3	4
17	14	72	2	5	3
18	7	33	3	1	0
19	17	66	5	2	4
20	12	60	3	7	3
21	19	105	2	4	9
22	2	12	1	0	0
23	16	68	4	6	8
24	8	24	4	1	5
25	5	24	2	3	2
26	10	61	3	2	4
27	12	46	4	2	4
28	10	55	3	2	2
29	12	46	4	2	7
30	17	104	3	1	15
31	8	19	3	3	3
32	4	19	1	0	0
33	3	28	1	1	0
34	3	19	2	1	0
35	10	56	3	4	3
36	5	29	2	0	2
37	17	55	5	6	10
38	8	44	1	0	1
39	8	30	2	2	3
40	13	56	5	3	7
41	19	78	8	4	3

References

Aikawa, Takako, Lee Schwartz, Ronit King, Monica Corston-Oliver and Carmen Lozano. 2007. "Impact of Controlled Language on Translation Quality and Post-Editing in a Statistical Machine Translation Environment." In *Proceedings of Machine Translation Summit Xi, 10-14 September 2007, Copenhagen, Denmark*, edited by Bente Maegaard, 1-7.

Alves, Fabio, Adriana Pagano and Igor da Silva. 2010. "A New Window on Translators' Cognitive Activity." In *Methodology, Technology and Innovation in Translation Process Research,* edited by Inger Mees, Fabio Alves and Susanne Göpferich, 267-291. Copenhagen: Samfundslitteratur.

Anderson, John R. 2005. *Cognitive Psychology and Its Implications*. 6th ed. New York: Worth Publishers.

Asia Online. 2012. "Fair Compensation for Post-Editing Machine Translation." Accessed 26 June 2014. http://www.asiaonline.net/EN/ Resources/Articles/FairCompensationForMTPostEditing.aspx.

Aziz, Wilker, Sheila Castilho and Lucia Specia. 2012. "PET: A Tool for Post-Editing and Assessing Machine Translation." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, 21-27 May 2012, Istanbul, Turkey,* edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis, 3982-3987. European Language Resources Association.

Aziz, Wilker, Maarit Koponen and Lucia Specia. 2014. "Sub-Sentence Level Analysis of Machine Translation Post-Editing Effort." In *Post-Editing of Machine Translation: Processes and Applications,* edited by Sharon O'Brien, Laura W. Balling, Michael Carl, Michel Simard and Lucia Specia, 170-199. Newcastle upon Tyne: Cambridge Scholars Publishing.

Baayen, R. Harald. 2008. *Analysing Linguistic Data: A Practical Introduction to Statistics Using R*. New York, USA and Cambridge, UK: Cambridge University Press.

Baayen, R. Harald, Douglas J. Davidson and Douglas M. Bates. 2008. “Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items.” *Journal of Memory and Language* 59 (4):390-412.

Babych, Bogdan and Anthony Hartley. 2008. “Sensitivity of Automated MT Evaluation Metrics on Higher Quality MT Output: BLEU vs Task-Based Evaluation Methods.” In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, 28-29-30 May 2008, Marrakech, Morocco,* edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis and Daniel Tapias, 2133-2136. European Language Resources Association.

Baddeley, Alan. 1996. “Exploring the Central Executive.” *The Quarterly Journal of Experimental Psychology Section A* 49 (1):5-28. doi: 10.1080/713755608.

Baddeley, Alan. 1999. *Essentials of Human Memory*. Hove: Psychology Press.

Baddeley, Alan. 2000. “The Episodic Buffer: A New Component of Working Memory?” *Trends in Cognitive Sciences* 4 (11):417-423. doi: http://dx.doi.org/10.1016/S1364-6613(00)01538-2.

Baddeley, Alan. 2007. *Working Memory, Thought, and Action*. New York, USA and Oxford, UK: Oxford University Press.

Baddeley, Alan. 2009a. “Short-Term Memory.” In *Memory*, edited by Alan Baddeley, Michael W. Eysenck and Michael C. Anderson, 19-40. Hove and New York: Psychology Press.

Baddeley, Alan. 2009b. “Working Memory.” In *Memory*, edited by Alan Baddeley, Michael W. Eysenck and Michael C. Anderson, 41-68. Hove and New York: Psychology Press.

Baddeley, Alan. 2009c. “What Is Memory?” In *Memory*, edited by Alan Baddeley, Michael W. Eysenck and Michael C. Anderson, 1-17. Hove and New York: Psychology Press.

Baddeley, Alan and Graham Hitch. 1974. “Working Memory.” In *The Psychology of Learning and Motivation: Advances in Research and Theory,* edited by Gordon A. Bower, 47-89. New York: Academic Press.

Baddeley, Alan, Giuseppe Vallar and Barbara Wilson. 1987. “Sentence Comprehension and Phonological Memory: Some Neuropsychological Evidence.” In *Attention and Performance Xii: The Psychology of Reading*, edited by Max Coltheart, 509-529. London: Lawrence Erlbaum Associates.

Balling, Laura W. 2008. “A Brief Introduction to Regression Designs and Mixed-Effects Modelling by a Recent Convert.” In *Looking at Eyes. Eye-Tracking Studies of Reading and Translation Processing,* edited by Susanne Göpferich, Arnt L. Jakobsen and Inger Mees, 175-192. Copenhagen: Samfundslitteratur.

Balling, Laura W. and Harald Baayen. 2008. “Morphological Effects in Auditory Word Recognition: Evidence from Danish.” *Language and Cognitive Processes* 23 (7-8):1159-1190. doi: 10.1080/01690960802201010.

Banerjee, Satanjeev and Alon Lavie. 2005. “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.” In *Proceedings of the ACL-05 Workshop, Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 29 June 2005, Ann Arbor, USA,* edited by Jade Goldstein, Alon Lavie, Chin-Yew Lin and Clare Voss, 65-72. Association for Computational Linguistics.

Bates, Douglas, Martin Maechler and Ben Bolker. 2012. Lme4: Linear Mixed-Effects Models Using S4 Classes. R Package Version 0.999999-0.

Bernardini, Silvia. 2001. “Think-Aloud Protocols in Translation Research.” *Target* 13 (2):241-263.

Bernth, Arendse and Claudia Gdaniec. 2002. “MTranslatability.” *Machine Translation* 16 (3):175-218.

Björnsson, Carl H. 1968. *Läsbarhet*. Stockholm: Liber.

Blain, Frédéric, Jean Senellart, Holger Schwenk, Mirko Plitt and Johann Roturier. 2011. “Qualitative Analysis of Post-Editing for High Quality Machine Translation.” In *Proceedings of the 13th Machine Translation Summit, 19-23 September 2011, Xiamen, China,* edited by Hiromi Nakaiwa, 164-171. Asia-Pacific Association for Machine Translation.

Bland, J. Martin and Douglas G. Altman. 1995. “Calculating Correlation Coefficients with Repeated Observations: Part 2—Correlation between Subjects.” *BMJ* 310:633. doi: 10.1136/bmj.310.6980.633.

Blatz, John, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis and Nicola Ueffing. 2004. “Confidence Estimation for Machine Translation.” In *Proceedings of the 20th International Conference on Computational Linguistics, 23-27 August 2004, Geneva, Switzerland,* 315-321. Association for Computational Linguistics.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia. 2013. “Findings of the 2013 Workshop on Statistical Machine Translation.” In *Proceedings of the Eighth Workshop on Statistical Machine Translation, 8-9 August 2013, Sofia, Bulgaria*, 1-44.

Brener, Roy. 1940. “An Experimental Investigation of Memory Span.” *Journal of Experimental Psychology* 26 (5):467-482.

Cain, Brad. 2007. A Review of the Mental Workload Literature. DTIC Document.

Callison-Burch, Chris and Miles Osborne. 2006. “Re-Evaluating the Role of BLEU in Machine Translation Research.” In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 3-7 April 2006, Trento, Italy*, 249-256.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz and Josh Schroeder. 2007. “(Meta-) Evaluation of Machine Translation.” In *Proceedings of the Second Workshop on Statistical Machine Translation, 23 June 2007, Prague, Czech Republic*, 136-158.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia. 2012. “Findings of the 2012 Workshop on Statistical Machine Translation.” In *Proceedings of the 7th Workshop on Statistical Machine Translation, 7-8 June 2012 Montréal, Canada*, 10–51. Association for Computational Linguistics.

Calude, Andreea S. 2002. Machine Translation of Various Text Genres. Paper presented at the *7th Language and Society Conference of the New Zealand Linguistic Society*. Hamilton, New Zealand.

Canek, David. 2011. “Wanted: A Fair and Simple Compensation Scheme for MT Post-Editing.” Accessed 14 August 2015. http://kv-emptypages.blogspot.co.uk/2011/11/wanted-fair-and-simple-compensation.html

Caplan, David and Gloria Waters. 2003. “The Relationship between Age, Processing Speed, Working Memory Capacity, and Language Comprehension.” *Memory* 13 (3-4):403-413. doi: 10.1080/09658210344000459.

Carl, Michael. 2012. “Translog-II: A Program for Recording User Activity Data for Empirical Reading and Writing Research.” In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, 21-27 May 2012*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet

Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis, 4108-4112. European Language Resources Association.

Carl, Michael, Martin Kay and Kristian T.H. Jensen. 2010. "Long Distance Revisions in Drafting and Post-Editing." In *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics, 21–27 March 2010, Iaşi, Romania*, 193-204.

Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt and Arnt L. Jakobsen. 2011. "The Process of Post-Editing: A Pilot Study." In *Human-Machine Interaction in Translation*, edited by Bernardette Sharp, Michael Zock, Michael Carl and Arnt L. Jakobsen, 131-142. Copenhagen: Samfundslitteratur.

Catling, Jonathan and Jonathan Ling. 2012. *Cognitive Psychology*. Harlow and New York: Prentice Hall.

Chow, Michael and Andrew R. A. Conway. 2015. "The Scope and Control of Attention: Sources of Variance in Working Memory Capacity." *Memory & Cognition* 43(3):325-339. doi: 10.3758/s13421-014-0496-9.

Christensen, Rune H. B. 2010. Ordinal—Regression Models for Ordinal Data. R Package Version 22.

Cowan, Nelson. 1999. "An Embedded-Process Model of Working Memory." In *Models of Working Memory*, edited by Miyake Akira and Priti Shah, 62-101. Cambridge: Cambridge University Press.

Cowan, Nelson. 2005. *Working Memory Capacity*. Hove and New York: Psychology Press.

Cowan, Nelson. 2010. "The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why?" *Current directions in psychological science* 19 (1):51-57. doi: 10.1177/0963721409359277.

Cowan, Nelson, Emily M. Elliott, J. Scott Saults, Candice C. Morey, Sam Mattox, Anna Hismjatullina and Andrew R. A. Conway. 2005. "On the Capacity of Attention: Its Estimation and Its Role in Working Memory and Cognitive Aptitudes." *Cognitive Psychology* 51 (1):42-100. doi: 10.1016/j.cogpsych.2004.12.001.

Creswell, John W. 2009. *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. 3rd ed. London: Sage.

Daneman, Meredyth and Patricia A. Carpenter. 1980. "Individual Differences in Working Memory and Reading." *Journal of Verbal Learning and Verbal Behaviour* 19 (4):450-466. doi: 10.1016/S0022-5371(80)90312-6.

de Almeida, Giselle. 2013. *Translating the Post-Editor: An Investigation on Post-Editing Changes and Correlations with Professional Experience*. PhD Thesis, School of Applied Language and Intercultural Studies, Dublin City University.

de Almeida, Giselle and Sharon O'Brien. 2010. "Analysing Post-Editing Performance: Correlations with Years of Translation Experience." In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation, 27-28 May 2010, St. Raphaël, France,* edited by François Yvon and Viggo Hansen.

de Groot, Annette M. B. 2000. "A Complex-Skill Approach to Translating and Interpreting." In *Tapping and Mapping the Processes of Translation and Interpreting: Outlooks on Empirical Research*, edited by Sonja Tirkkonen-Condit and Riitta Jääskeläinen, 53-68. Amsterdam: Benjamins.

Denkowski, Michael and Alon Lavie. 2010. "METEOR-Next and the METEOR Paraphrase Tables: Improved Evaluation Support for Five Target Languages." In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, 15-16 July 2010, Uppsala, Sweden*, 339-342. Association for Computational Linguistics.

Denkowski, Michael and Alon Lavie. 2011. "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems." In *Proceedings of the Sixth Workshop on Statistical Machine Translation, 30-31 July 2011, Edinburgh, UK*, 85-91. Association for Computational Linguistics.

Doherty, Stephen, Sharon O'Brien and Michael Carl. 2010. "Eye Tracking as an MT Evaluation Technique." *Machine Translation* 24 (1):1-13. doi: 10.1007/s10590-010-9070-9.

Dragsted, Barbara and Inge G. Hansen. 2008. "Comprehension and Production in Translation: A Pilot Study on Segmentation and the Coordination of Reading and Writing Processes." In *Looking at Eye: Eye-Tracking Studies of Reading and Translation Processing,* edited by Susanne Göpferich, Arnt L. Jakobsen and Inger Mees, 9-29. Copenhagen: Samfundslitteratur.

Dragsted, Barbara and Michael Carl. 2013. "Towards a Classification of Translation Styles Based on Eye-Tracking and Keylogging Data." *Journal of Writing Research* 5 (1):133-158.

DuBay, William H. 2004. *The Principles of Readability*. Costa Mesa: Impact Information. Accessed 17 June 2015. http://www.impact-information.com/impactinfo/readability02.pdf

Duchowski, Andrew T. 2007. *Eye Tracking Methodology: Theory and Practice*. London: Springer-Verlag.

Ericsson, K. Anders and Herbert A. Simon. 1980. “Verbal Reports as Data.” *Psychological Review* 87 (3):215.

Ericsson, K. Anders and Herbert A. Simon. 1984. *Protocol Analysis: Verbal Reports as Data*. Cambridge and London: The Massachusetts Institute of Technology Press.

Ericsson, K. Anders and Walter Kintsch. 1995. “Long-Term Working Memory.” *Psychological Review* 102 (2):211–245.

Flesch, Rudolph. 1948. “A New Readability Yardstick.” *Journal of Applied Psychology* 32 (3):221-233. doi: 10.1037/h0057532.

Francis, Mardean and Deborah McCutchen. 1994. “Strategy Differences in Revising between Skilled and Less Skilled Writers.” *Paper presented at the Annual Meeting of the American Educational Research Association, April 1994, New Orleans, USA.*

FreePressRelease. 2011. “As Content Volume Explodes, Machine Translation Becomes an Inevitable Part of Global Content Strategy.” Accessed 29 October 2012. http://www.freepressrelease.eu/?p=42044.

Gaspari, Federico, Antonio Toral, Sudip Kumar Naskar, Declan Groves and Andy Way. 2014. “Perception Vs Reality: Measuring Machine Translation Post-Editing Productivity.” In *Proceedings of the Third Workshop on Post-Editing Technology and Practice, the 11th Conference of the Association for Machine Translation in the Americas, 22-26 October 2014, Vancouver, Canada*, edited by Sharon O’Brien, Michel Simard and Lucia Specia, 60-72.

Gile, Daniel. 1995. *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam: John Benjamins.

Gopher, Daniel and Rolf Braune. 1984. “On the Psychophysics of Workload: Why Bother with Subjective Measures?” *Human Factors* 26:519-532.

Gopher, Daniel and Emanuel Donchin. 1986. “Workload - an Examination of a Concept.” In *Handbook of Perception and Human Performance*, edited by Kenneth R. Boff, Lloyd Kaufman and James P. Thomas, 1-49. New York: John Wiley and Sons.

Gouadec, Daniel. 2010. *Translation as a Profession*. Corrected ed. Amsterdam and Philadelphia: John Benjamins.

Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse and Zhiqiang Cai. 2004. “Coh-Metrix: Analysis of Text on Cohesion and Language.” *Behavior Research Methods, Instruments, & Computers* 36 (2):193-202.

Graesser, Arthur C. and Danielle S. McNamara. 2011. “Computational Analyses of Multilevel Discourse Comprehension.” *Topics in Cognitive Science* 3 (2):371-398. doi: 10.1111/j.1756-8765.2010.01081.x.

Green, John M. and Rebecca Oxford. 1995. “A Closer Look at Learning Strategies, L2 Proficiency, and Gender.” *TESOL Quarterly* 29 (2):261-297.

Green, Spence, Marie-Catherine de Marneffe, John Bauer and Christopher D. Manning. 2011. “Multiword Expression Identification with Tree Substitution Grammars: A Parsing Tour De Force with French.” In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 27–31 July 2011, Edinburgh, UK*, 725-735.

Green, Spence, Jeffrey Heer and Christopher D. Manning. 2013. “The Efficacy of Human Post-Editing for Language Translation.” In *ACM Human Factors in Computing Systems (Chi), 27 April - 2 May 2013, Paris, France*, 439-448.

Green, Spence, Jason Chuang, Jeffrey Heer and Christopher D. Manning. 2014. “Predictive Translation Memory: A Mixed-Initiative System for Human Language Translation.” In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, 05-08 October 2014, Honolulu, USA*, 177-187. Association for Computing Machinery.

Guerberof, Ana. 2013. “What do professional translators think about post-editing?” The *Journal of Specialised Translation* 13:75-95.

Guerberof, Ana. 2014. “The Role of Professional Experience in Post-Editing from a Quality and Productivity Perspective.” In *Post-Editing of Machine Translation: Processes and Applications,* edited by Sharon O’Brien, Laura W. Balling, Michael Carl, Michel Simard and Lucia Specia, 51-76. Newcastle upon Tyne: Cambridge Scholars Publishing.

Guzmán, Rafael. 2007. “Manual MT Post-Editing: If It’s Not Broken, Don’t Fix It! “ *Translation Journal* 11 (4). Accessed 26 June 2014. http://www.translationjournal.net/journal/42mt.htm.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten. 2009. “The Weka Data Mining Software: An Update.” *SIGKDD Explorations* 11 (1):10-18.

Hart, Sandra G. and Lowell E. Staveland. 1988. “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research.” In *Human Mental Workload,* edited by A. Hancock Peter and Meshkati Najmedin, 139-183. Amsterdam: North-Holland.

Hogarth, Robin M. 2012. “Deciding Analytically or Trusting Your Intuition? The Advantages and Disadvantages of Analytic and Intuitive Thought.” In *The Routines of Decision Making*, edited by Tilmann Betsch and Susanne Haberstroh, 67-82. Hove and New York: Psychology Press.

Hvelplund, Kristian T. 2011. *Allocation of Cognitive Resources in Translation: An Eye-Tracking and Key-Logging Study*. PhD Thesis, The Doctoral School of Language, Law, informatics, Operations Management and Culture, Copenhagen Business School.

Inhoff, Albrehct W. and Ralph Radach. 1998. “Definition and Computation of Oculomotor Measures in the Study of Cognitive Processes.” In *Eye Guidance in Reading and Scene Perception*, edited by Geoffrey Underwood, 29-54. Oxford: Elsevier.

Irwin, David E. 1998. “Lexical Processing During Saccadic Eye Movements.” *Cognitive Psychology* 36 (1):1-27. doi: http://dx.doi.org/10.1006/cogp.1998.0682.

Jääskeläinen, Riitta. 1996. “Hard Work Will Bear Beautiful Fruit: A Comparison of Two Think-Aloud Protocol Studies.” *Meta* XLI (1):60-74.

Jääskeläinen, Riitta. 2002. “Think-Aloud Protocol Studies into Translation: An Annotated Bibliography.” *Target* 14 (1):107-136.

Jääskeläinen, Riitta. 2010. “Are All Professionals Experts? Definitions of Expertise and Reinterpretation of Research Evidence in Process Studies.” In *Translation and Cognition*, edited by Gregory Shreve and Erik Angelone, 213–227. Amsterdam and Philadelphia: John Benjamins.

Jääskeläinen, Riitta and Sonja Tirkkonen-Condit. 1991. “Automatised Processes in Professional vs. Non-Professional Translation: A Think-Aloud Protocol Study.” In *Empirical Research in Translation and Intercultural Studies - Selected Papers of the Transif Seminar, Savonlinna 1988,* edited by Sonja Tirkkonen-Condit, 265-269. Tübingen: Gunter Narr Verlag.

Jakobsen, Arnt L. 2003. “Effects of Think-Aloud on Translation Speed, Revision and Segmentation.” In *Triangulating Translation. Perspectives in Process Oriented*

Research, edited by Fabio Alves, 69-95. Amsterdam and Philadelphia: John Benjamins.

Jakobsen, Arnt L. and Kristian T. H. Jensen. 2008. “Eye Movement Behaviour across Four Different Types of Reading Task.” In *Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing*, edited by Susanne Göpferich, Arnt L. Jakobsen and Inger Mees, 103-124. Copenhagen: Samfundslitteratur.

James, William. 1890. *Principles of Psychology*. New York: Holt.

Jensen, Christian. 2008. “Assessing Eye-Tracking Accuracy in Translation Studies.” In *Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing*, edited by Susanne Göpferich, Arnt L. Jakobsen and Inger Mees, 157-174. Copenhagen: Samfundslitteratur.

Jensen, Kristian T.H., Annette C. Sjørup and Laura W. Balling. 2010. “Effects of L1 Syntax on L2 Translation.” In *Methodology, Technology and Innovation in Translation Process Research*, edited by Inger Mees, Fabio Alves and Susanne Göpferich, 319-339. Copenhagen: Samfundslitteratur.

Jex, Henry R. 1988. “Measuring Mental Workload: Problems, Progress, and Promises.” In *Human Mental Workload*, edited by Peter Hancock, A. and Najmedin Meshkati, 5-39. Amsterdam: North-Holland.

Jones, Francis R. 2011. *Poetry Translating as Expert Action: Processes, Priorities and Networks*. Amsterdam and Philadelphia: John Benjamins.

Jones, Glyn. 2000. “Compiling French Word Frequency Lists for the VAT: A Feasibility Study.” Accessed 20 December 2013. http://www.lextutor.ca/vp/fr/.

Just, Marcel A. and Patricia A. Carpenter. 1980. “A Theory of Reading: From Eye Fixation to Comprehension.” *Psychological Review* 87:329-354.

Kahneman, Daniel. 1973. *Attention and Effort*. Englewood Cliffs, N.J.: Prentice-Hall.

Kandel, Liliane and Abraham Moles. 1958. “Application de l’indice de Flesch à la langue française.” *Cahiers Etudes de Radio-Télévision* 19:253-274.

Kirschner, Paul A. 2002. “Cognitive Load Theory: Implications of Cognitive Load Theory on the Design of Learning.” *Learning and Instruction* 12:1-10.

Koby, Geoffrey S. and Gertrud G. Champe. 2013. “Welcome to the Real World: Professional-Level Translator Certification.” *Translation and Interpreting* 5 (1):156-173. doi: ti.105201.2013.a09.

Koponen, Maarit. 2012. “Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations.” In *Proceedings of the Seventh Workshop on Statistical Machine Translation, 7-8 June 2012, Montréal, Canada*, 181-190.

Koponen, Maarit, Wilker Aziz, Luciana Ramos and Lucia Specia. 2012. “Post-Editing Time as a Measure of Cognitive Effort.” In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012), 28 October 2012, San Diego, USA,* edited by Sharon O’Brien, Michel Simard and Lucia Specia.

Krings, Hans P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Edited by Geoffrey S. Koby. Kent: Kent State University Press.

Krippendorff, Klaus. 2011. “Computing Krippendorff’s Alpha Reliability.” Accessed 17 June 2015. http://repository.upenn.edu/asc_papers/43.

Kuznetsova, Alexandra, Per B. Brockhoff and Rune H. B. Christensen. 2013. Lmertest: Tests for Random and Fixed Effects for Linear Mixed Effect Models (Lmer Objects of Lme4 Package). R Package Version 2.0-20.

Lacruz, Isabel and Gregory Shreve. 2014. “Pauses and Cognitive Effort in Post-Editing.” In *Post-Editing of Machine Translation: Processes and Applications*, edited by Sharon O’Brien, Laura W. Balling, Michael Carl, Michel Simard and Lucia Specia, 246-272. Newcastle upon Tyne: Cambridge Scholars Publishing.

Lacruz, Isabel, Michael Denkowski and Alon Lavie. 2014. “Cognitive Demand and Cognitive Effort in Post-Editing.” In *Proceedings of the Third Workshop on Post-Editing Technology and Practice, 11th Conference of the Association for Machine Translation in the Americas, 22-26 October 2014, Vancouver, Canada*, edited by Sharon O’Brien, Michel Simard and Lucia Specia, 73-84.

Landis, J. Richard and Gary G. Koch. 1977. “The Measurement of Observer Agreement for Categorical Data.” *Biometrics* 33 (1):159-174.

Larigauderie, Pascale, Daniel Gaonac’h and Natasha Lacroix. 1998. “Working Memory and Error Detection in Texts: What Are the Roles of the Central Executive and the Phonological Loop?” *Applied Cognitive Psychology* 12 (5):505-527.

Linguistic Data Consortium. 2005. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations. Revision 1.5.

Li, Defeng. 2004. “Trustworthiness of Think-Aloud Protocols in the Study of Translation Processes.” *International Journal of Applied Linguistics* 14 (3):301-313. doi: 10.1111/j.1473-4192.2004.00067.x.

Lin, Chin-Yew and Franz Josef Och. 2004. “ORANGE: A Method for Evaluating Automatic Evaluation Metrics for Machine Translation.” In *Proceedings of the*

20th International Conference on Computational Linguistics, 23-27 August 2004, Geneva, Switzerland. Association for Computational Linguistics.

Longo, Luca and Stephen Barrett. 2010. “A Computational Analysis of Cognitive Effort.” In *Proceedings of the Second International Conference on Intelligent Information and Database Systems: Part II, Lecture Notes in Computer Science*, edited by Ngoc Thanh Nguyen, Mann Thanh Le and Jerzy Świątek, 65–74. Berlin and Heidelberg: Springer-Verlag.

Lopez, Adam. 2012. “Putting Human Assessments of Machine Translation Systems in Order.” In *Proceedings of the 7th Workshop on Statistical Machine Translation, 7-8 June 2012, Montréal, Canada*, 1-9. Association for Computational Linguistics.

Machácek, Matouš and Ondrej Bojar. 2013. “Results of the WMT13 Metrics Shared Task.” In *Proceedings of the Eighth Workshop on Statistical Machine Translation, 8-9 August 2013, Sofia, Bulgaria*, 45–51. Association for Computational Linguistics.

MacQueen, James B. 1967. “Some Methods for Classification and Analysis of Multivariate Observations.” In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 21 June - 18 July 1965 and 27 December 1965 - 7 January 1966, Berkeley, USA*, edited by Lucien M. Le Cam and Jerzy Neyman, 281-297. Berkeley: University of California Press.

Martín, Ricardo Muñoz. 2010. “On Paradigms and Cognitive Translatology.” In *Translation and Cognition*, edited by Gregory Shreve and Erik Angelone, 169-187. Amsterdam and Philadelphia: John Benjamins.

May, James G., Robert S. Kennedy, Mary C. Williams, William P. Dunlap and Julie R. Brannan. 1990. “Eye Movement Indices of Mental Workload.” *Acta Psychologica* 75 (1):75-89.

McCutchen, Deborah. 1996. “A Capacity Theory of Writing: Working Memory in Composition.” *Educational Psychology Review* 8:299-325. doi: 10.1007/BF01464076.

Meara, Paul and Barbara Buxton. 1987. “An Alternative to Multiple Choice Vocabulary Tests.” *Language Testing* 4 (2):142-154.

Melamed, I. Dan, Ryan Green and Joseph P. Turian. 2003. “Precision and Recall of Machine Translation.” In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human*

Language Technology, 31 May - 1 June 2003, Edmonton, Canada: Companion Volume of the Proceedings of HLT-NAALC 2003--Short Papers - Volume 2, 61-63. Association for Computational Linguistics.

Miller, George A. 1956. “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information.” *The Psychological Review* 63:81-97.

Miller, George A. 1995. “WordNet: A Lexical Database for English”. *Communications of the ACM* 38 (11): 39-41.

Mitchell, Linda, Johann Roturier and Sharon O’Brien. 2013. “Community-Based Post-Editing of Machine-Translated Content: Monolingual vs. Bilingual.” In *Proceedings of the MT Summit XIV, Workshop on Post-Editing Technology and Practice, 2 September 2013, Nice, France,* edited by Sharon O’Brien, Michel Simard and Lucia Specia, 35-43.

Moray, Neville. 1967. “Where Is Capacity Limited? A Survey and a Model.” *Acta Psychologica* 27 (0):84-92. doi: http://dx.doi.org/10.1016/0001-6918(67)90048-0.

Mossop, Brian. 2007. *Revising and Editing for Translators*. Manchester and Kinderhook: St. Jerome.

Nakov, Preslav, Francisco Guzmán and Stephan Vogel. 2012. “Optimizing for Sentence-Level BLEU+ 1 Yields Short Translations.” In *Proceedings of COLING 2012: Technical Papers, 8-15 December 2012, Mumbai, India*, edited by Martin Kay and Christian Boitet, 1979-1994. Mumbai: COLING 2012 Organizing Committee.

Newmark, Peter. 1988. *A Textbook of Translation*. Englewood Cliffs, N.J.: Prentice-Hall.

Norman, Donald A. and Tim Shallice. 1986. “Attention to Action: Willed and Automatic Control of Behaviour.” In *Consciousness and Self-Regulation: Advances in Research and Theory,* edited by Richard J. Davidson, Gary E. Schwarts and David Shapiro, 1-18. New York: Plenum Press.

O’Brien, Sharon. 2004. “Machine Translatability and Post-Editing Effort: How Do They Relate?” In *Translating and the Computer 26, November 2004, London, UK*. London: Aslib.

O’Brien, Sharon. 2005. “Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability.” *Machine Translation* 19 (1):37-58. doi: 10.1007/s10590-005-2467-1.

O’Brien, Sharon. 2006. “Controlled Language and Post-Editing” *MultiLingual* October/November issue:17-19.

O’Brien, Sharon. 2010. “Eye Tracking in Translation-Process Research: Methodological Challenges and Solutions.” In *Methodology, Technology and Innovation in Translation Process Research,* edited by Inger Mees, Fabio Alves and Susanne Göpferich, 251-266. Copenhagen: Samfundslitteratur.

O’Brien, Sharon. 2011. “Towards Predicting Post-Editing Productivity.” *Machine Translation* 25 (3):197-215. doi: 10.1007/s10590-011-9096-7.

O’Brien, Sharon, Rahzeb Choudhury, Jaap van der Meer and Nora A. Monasterio. 2011. TAUS Report - Dynamic Quality Evaluation Framework. De Rijp, The Netherlands: TAUS.

O’Donnell, Robert D. and F. Thomas Eggemeier. 1986. “Workload Assessment Methodology.” In *Handbook of Perception and Human Performance*, edited by Kenneth R. Boff, Lloyd Kaufman and James P. Thomas, 42-1-42-49. New York: John Wiley and Sons.

Offersgaard, Lene, Claus Povlsen, Lisbeth K. Almsten and Bente Maegaard. 2008. “Domain Specific MT in Use.” In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation, 22-23 September 2008, Hamburg, Germany*, edited by John Hutchins and Walther v. Hahn, 150-159. Hamburg: Hamburger Informatik Technologie-Center e.V.

Paas, Fred. 1992. “Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach.” *Journal of Educational Psychology* 84 (4):429-434.

Paas, Fred and Jeroen J. G. van Merriënboer. 1994. “Variability of Worked Examples and Transfer of Geometrical Problem-Solving Skills: A Cognitive-Load Approach.” *Journal of Educational Psychology* 86 (1):122-133.

Paas, Fred, Juhani E. Tuovinen, Huib Tabbers and Pascal W. M. van Gerven. 2003. “Cognitive Load Measurement as a Means to Advance Cognitive Load Theory.” *Educational Psychologist* 38 (1):63-71.

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. “BLEU: A Method for Automatic Evaluation of Machine Translation.” In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 7-12 July 2002, Philadelphia USA*, 311-318. Association for Computational Linguistics.

Plitt, Mirko and François Masselot. 2010. "A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localization Context." *The Prague Bulletin of Mathematical Linguistics* 93:7-16.

Posner, Michael I. 1980. "Orienting of Attention." *Quarterly Journal of Experimental Psychology* 32:3-25.

Radach, Ralph, Alan Kennedy and Keith Rayner. 2004. *Eye Movements and Information Processing During Reading*. Hove and New York: Psychology Press.

Rayner, Keith. 1998. "Eye Movements in Reading and Information Processing: 20 Years of Research." *Psychological Bulletin* 124 (3):372-422.

Read, John. 2007. "Second Language Vocabulary Assessment: Current Practices and New Directions." *International Journal of English Studies* 7 (2):105-126.

Redick, Thomas S., James M. Broadway, Matt E. Meier, Princy S. Kuriakose, Nash Unsworth, Michael J. Kane and Randall W. Engle. 2012. "Measuring Working Memory Capacity with Automated Complex Span Tasks." *European Journal of Psychological Assessment* 28 (3):164-171.

Reid, Gary B. and Thomas E. Nygren. 1988. "The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload." In *Human Mental Workload*, edited by A. Hancock Peter and Meshkati Najmedin, 185-218. Amsterdam: North-Holland.

Roodenrys, Kylie, Shirley Agostinho, Steven Roodenrys and Paul Chandler. 2012. "Managing One's Own Cognitive Load When Evidence of Split Attention Is Present." *Applied Cognitive Psychology* 26 (6):878-886. doi: 10.1002/acp.2889.

Schneider, Walter and Richard M. Shiffrin. 1977a. "Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending and a General Theory." *Psychological Review* 84 (2):127-190.

Schneider, Walter and Richard M. Shiffrin. 1977b. "Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention." *Psychological Review* 84 (1):1-66.

Schoonen, Rob, Patrick Snellings, Marie Stevenson and Amos van Gelderen. 2009. "Towards a Blueprint of the Foreign Language Writer: The Linguistic and Cognitive Demands of Foreign Language Writing." In *Writing in Foreign Language Contexts: Learning, Teaching, and Research*, edited by Rosa M. Manchón, 77-101. Bristol, Buffalo and Toronto: Multilingual Matters.

Shih, Claire Yi-Yi. 2006. *Translators' Revision Processes: Global Revision Approaches and Strategic Revision Behaviours*. PhD thesis, School of Modern Languages, Newcastle University.

Shute, Valerie J. 1991. "Who Is Likely to Acquire Programming Skills?" *Journal of Educational Computing Research* 7 (1):1-24. doi: 10.2190/VQJD-T1YD-5WVB-RYPJ.

Sirén, Seija and Kai Hakkarainen. 2002. "Expertise in Translation." *Across Languages and Cultures* 3 (1):71-82.

Sjørup, Annette C. 2011. "Cognitive Effort in Metaphor Translation: An Eye-Tracking Study." In *Cognitive Explorations of Translation*, edited by Sharon O'Brien, 197-214. London: Continuum International.

Snijders, Tom A. B. and Roel Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul. 2006. "A Study of Translation Edit Rate with Targeted Human Annotation." In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas, Visions for the Future of Machine Translation, 8-12 August 2006, Cambridge, USA*, 223-231.

Snover, Matthew G., Nitin Madnani, Bonnie Dorr and Richard Schwartz. 2009. "TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate." *Machine Translation* 23 (2-3):117-127.

Song, Xingyi, Trevor Cohn and Lucia Specia. 2013. "BLEU Deconstructed: Designing a Better MT Evaluation Metric." *International Journal of Computational Linguistics and Applications* 4 (2):29–44.

Specia, Lucia. 2011. "Exploiting Objective Annotations for Measuring Translation Post-Editing Effort." In *Proceedings of the 15th International Conference of the European Association for Machine Translation, 30-31 May 2011, Leuven, Belgium*, edited by Mikel L. Forcada, Ilse Depraetere and Vincent Vandeghinste, 73-80.

Specia, Lucia, Marco Turchi, Nicola Cancedda, Marc Dymetman and Nello Cristianini. 2009. "Estimating the Sentence-Level Quality of Machine Translation Systems." In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation, 14-15 May 2009, Barcelona, Spain*, 28-37. Barcelona: Universitat Politècnica de Catalunya.

Specia, Lucia, Dhwaj Raj and Marco Turchi. 2010. "Machine Translation Evaluation Versus Quality Estimation." *Machine Translation* 24 (1):39-50.

Specia, Lucia and Kashif Shah. 2013. Deliverable D2. 1.1 Quality Estimation Baseline Software.

Squire, Larry R. 1992. "Declarative and Nondeclarative Memory: Multiple Brain Systems Supporting Learning and Memory." *Journal of Cognitive Neuroscience* 4 (3):232-243. doi: 10.1162/jocn.1992.4.3.232.

Sun, Sanjun. 2011. "Think-Aloud Translation Process Research: Some Methodological Considerations." *Meta* LVI (4):928-951.

Sun, Sanjun and Gregory Shreve. 2014. "Measuring Translation Difficulty: An Empirical Study." *Target* 26 (1):98-127.

Sweller, John, Paul Ayres and Slava Kalyuga. 2011. *Cognitive Load Theory*. New York: Springer.

Tabbers, Huib K., Rob L. Martens and Jeroen J. G. van Merriënboer. 2004. "Multimedia Instructions and Cognitive Load Theory: Effects of Modality and Cueing." *British Journal of Educational Psychology* 74 (1):71-81. doi: 10.1348/000709904322848824.

Tatsumi, Midori. 2009. "Correlation between Automatic Evaluation Scores, Post-Editing Speed and Some Other Factors." In *Proceedings of the Twelfth Machine Translation Summit, 26–30 August 2009, Ottawa, Canada*, 332–339.

Tatsumi, Midori. 2010. *Post-Editing Machine Translated Text in a Commercial Setting: Observation and Statistical Analysis*. PhD Thesis, School of Applied Language and Intercultural Studies, Dublin City University.

Tatsumi, Midori and Johann Roturier. 2010. "Source Text Characteristics and Technical and Temporal Post-Editing Effort: What Is Their Relationship?" In *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry", 4 November 2010, Denver USA*, edited by Ventsislav Zhechev, 43-51.

TAUS. 2010. "TAUS Research - Postediting in Practice." Accessed 06 May 2015. http://taus-website-media.s3.amazonaws.com/images/stories/pdf/benchmarch-data-for-postediting-practices-globally.pdf.

TAUS. 2013a. "Adequacy/Fluency Guidelines." Accessed 17 June 2015. https://www.taus.net/think-tank/best-practices/evaluate-best-practices/adequacy-fluency-guidelines.

TAUS. 2013b. “Pricing Machine Translation Post-Editing Guidelines.” Accessed 24 June 2014. https://evaluation.taus.net/resources-c/pricing-machine-translation-post-editing-guidelines.

TAUS/CNGL. 2010. “MT Post-editing Guidelines.” Accessed 17 June 2015. https://www.taus.net/think-tank/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines.

Temnikova, Irina. 2010. “A Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment.” In *Proceedings of the 7th International Conference on Language Resources and Evaluation, 19-21 May 2010, Valetta, Malta*, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias, 3485-3490. European Language Resources Association.

Tirkkonen-Condit, Sonja. 2005. “The Monitor Model Revisited: Evidence from Process Research.” *Meta* L (2):405-414.

Tobii Technology. 2012. “Determining the Tobii I-VT Fixation Filter’s Default Values - Method Description and Results Discussion.” Accessed 20 December 2013. http://www.tobii.com/Global/Analysis/ Training/WhitePapers/Tobii_WhitePaper_DeterminingtheTobiiI-VTFixationFilter’sDefaultValues.pdf.

Toglia, Michael P. and William F. Battig. 1978. *Handbook of Semantic Word Norms*. New York: Lawrence Erlbaum.

Toutanova, Kristina, Dan Klein, Christopher D. Manning and Yoram Singer. 2003. “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network.” In *Proceedings of HLT-NAALC 2003, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 27 May - 1 June, Edmonton, Canada*, 173-180. Association for Computational Linguistics.

Traxler, Matthew J. 2012. *Introduction to Psycholinguistics: Understanding Language Science*. 1st ed. Chichester and Malden: Wiley-Blackwell.

Tulving, Endel. 1972. “Episodic and Semantic Memory.” In *Oganization of Memory,* edited by Endel Tulving, Wayne Donaldson and Gordon H. Bower, 381-403. New York and London: Academic Press.

Turian, Joseph P., Luke Shen and I. Dan Melamed. 2003. “Evaluation of Machine Translation and Its Evaluation.” In *Proceedings of the MT Summit IX, 23-27 September 2003, New Orleans, USA*, 386-393.

Tyler, Sherman W., Paula T. Hertel, Marvin C. McCallum and Henry C. Ellis. 1979. “Cognitive Effort and Memory.” *Journal of Experimental Psychology: Human Learning and Memory* 5 (6):607-617.

Underwood, Nancy and Bart Jongejan. 2001. “Translatability Checker: A Tool to Help Decide Whether to Use MT.” In *Proceedings of MT Summit VIII, Machine Translation in the Information Age, 18-22 September 2001, Santiago De Compostela, Spain*, edited by Bente Maegaard, 363-368.

Unsworth, Nash, Richard P. Heitz, Josef C. Schrock and Randall W. Engle. 2005. “An Automated Version of the Operation Span Task.” *Behavior Research Methods* 37 (3):498-505.

Unsworth, Nash, Thomas S. Redick, Richard P. Heitz, James M. Broadway and Randall W. Engle. 2009. “Complex Working Memory Span Tasks and Higher-Order Cognition: A Latent-Variable Analysis of the Relationship between Processing and Storage.” *Memory* 17 (6):635-654.

van Dijk, Teun A. and Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. Orlando: Academic Press.

van Gog, Tamara, Liesbeth Kester, Fleurie Nievelstein, Bas Giesbers and Fred Paas. 2009. “Uncovering Cognitive Processes: Different Techniques that Can Contribute to Cognitive Load Research and Instruction.” *Computers in Human Behavior* 25 (2):325-331. doi: http://dx.doi.org/10.1016/j.chb.2008.12.021.

Vieira, Lucas N. 2012. “PT-EN Collocation Equivalents and Machine Translation Evaluation.” *BULAG Natural Language Processing and Human Language Technology* 37:117-136.

Vieira, Lucas N. 2013. “An Evaluation of Tools for Post-Editing Research: The Current Picture and Further Needs.” In *Proceedings of MT Summit XIV Workshop on Post-Editing Technology and Practice, 2 September 2013, Nice, France*, edited by Sharon O’Brien, Simard Michel and Lucia Specia, 93–101.

Vieira, Lucas N. 2014. “Indices of Cognitive Effort in Machine Translation Post-Editing.” *Machine Translation* 28 (3-4):187-216. doi: 10.1007/s10590-014-9156-x.

Wickens, Christopher D. 1984. “Processing Resources in Attention.” In *Varieties of Attention*, edited by Raja Parasuraman and David R. Davies, 63-102. Orlando and London: Academic Press.

Wu, Shao-Chuan. 2010. *Assessing Simultaneous Interpreting: A Study on Test Reliability and Examiners' Assessment Behaviour*. PhD thesis, School of Modern Languages, Newcastle University.

Yeo, Gillian and Andrew Neal. 2008. "Subjective Cognitive Effort: A Model of States, Traits, and Time." *Journal of Applied Psychology* 93 (3):617-631.