Investigating the Effects of Controlled Language on the Reading and Comprehension of Machine Translated Texts: A Mixed-Methods Approach

Stephen Doherty

Thesis submitted for the degree of Doctor of Philosophy

School of Applied Language and Intercultural Studies
Dublin City University
January 2012

Supervisors:
Dr. Sharon O'Brien & Dr. Dorothy Kenny

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____ ID No.: _____

Date: _____

Abstract

This study investigates whether the use of controlled language (CL) improves the readability and comprehension of technical support documentation produced by a statistical machine translation system. Readability is operationalised here as the extent to which a text can be easily read in terms of formal linguistic elements; while comprehensibility is defined as how easily a text's content can be understood by the reader.

A biphasic mixed-methods triangulation approach is taken, in which a number of quantitative and qualitative evaluation methods are combined. These include: eye tracking, automatic evaluation metrics (AEMs), retrospective interviews, human evaluations, memory recall testing, and readability indices. A further aim of the research is to investigate what, if any, correlations exist between the various metrics used, and to explore the cognitive framework of the evaluation process.

The research finds that the use of CL input results in significantly higher scores for items recalled by participants, and for several of the eye tracking metrics: fixation count, fixation length, and regressions. However, the findings show slight insignificant increases for readability indices and human evaluations, and slight insignificant decreases for AEMs. Several significant correlations between the above metrics are identified as well as predictors of readability and comprehensibility.

Publications & Presentations from this Research Project

Doherty, S. 2011. Using Controlled Language to Improve Statistical Machine Translation. Dublin Innovation Festival. November 16th, 2011, Croke Park, Dublin, Ireland.

Doherty, S. 2011. Cognitive Explorations of Think-Aloud Protocols: Insights from Eye Tracking. School of Applied Language & Intercultural Studies Research Showcase, November 9th, 2011. Dublin City University, Ireland.

Doherty, S. 2011. Measuring the Impact of Controlled Language on Machine Translation via Readability and Comprehensibility. Computational Linguistics in the Netherlands (CLIN) 21, February 11th, 2011. Ghent, Belgium.

Doherty, S., O'Brien, S. & Carl, M. 2010. Eye tracking as an MT evaluation technique. Machine Translation, 24, 1. Springer.

Doherty, S., O'Brien, S. & Kenny, D. 2010. Improving MT Output with Controlled Language. Centre for Next Generation Localisation - Annual Science Foundation of Ireland Review. July 11th, 2010, The Helix, Dublin, Ireland.

Doherty, S. 2009. Readability and Controlled Language in User-Based Machine Translation Evaluation and Eye-Tracking. International Post-graduate Conference in Translation & Interpreting 5, 21-22 November 2009. Edinburgh, United Kingdom.

Doherty, S. & O'Brien, S. 2009. Can MT output be evaluated through eye tracking? MT Summit XII: Proceedings of the Twelfth Machine Translation Summit, Ottawa, Ontario, Canada, pp. 214-221.

Doherty, S., O'Brien, S. & Kenny, D. 2009. Can MT output be evaluated through eye tracking? Department of Humanities & Social Sciences Post-Graduate Showcase, October 17th, 2009. Dublin City University, Ireland.

Acknowledgements

First and foremost, I must express the greatest gratitude to my supervisors Drs. Sharon O'Brien and Dorothy Kenny for their encouragement, challenges, and for always having time to talk, and listen, to me. Even before the research for the current study began, they were generous with their time and energy, and provided me with endless opportunities and support.

Thanks are due to Prof. Andy Way for his encouraging words and ideas that always came when needed most. Drs. Johann Roturier and Fred Hollowood were invaluable and generous with their time, resources, and patience, which were pivotal to the development of this project. Additionally, Dr. Sudip Kumar Naskar provided kindness and technical expertise throughout. I'd also like to mention the contributions of Dr. Minako O'Hagan and Prof. Jenny Williams for creating a wonderful environment where interesting research and teaching opportunities were made available to me, and where my participation and opinions always felt valued.

I'm grateful to Joss for his friendship over recent years and the thrifty coffee breaks. The time spent on this project would have not been as enjoyable without the friends I made in DCU: Midori, Yanli, Yanjun, Michele, Manuela, Yvette, Nora, and my colleagues in the Centre for Next Generation Localisation, the Centre for Translation and Textual Studies, and the School of Applied Language and Intercultural Studies.

Great thanks are also due to the funder of this project, the Science Foundation of Ireland, which provided a generous scholarship over the last 3 years without which none of this work would be possible. I'd also like to express my thanks to the participants in the studies, and the examiners for sharing their valuable time and experience with me.

Finally, I also wish to thank my family and friends for their continued support throughout the years of my education, even when it seemed at times to be endless.

Table of Contents

Part II: Methodological Considerations

Chapter Three: Methodology

Part III: Data Analysis

Chapter Six: Results & Discussion

List of Figures

List of Tables

List of Abbreviations

AEM - Automatic Evaluation Metric

BLEU - Bilingual Evaluation Understudy

CL - Controlled Language

EBMT – Example-Based Machine Translation

GTM - General Text Matcher

LTM – Long-Term Memory

MT - Machine Translation

RBMT - Rule-Based Machine Translation

SD - Standard Deviation

SMT - Statistical Machine Translation

ST - Source Text

STM – Short-Term Memory

TAP – Think-Aloud Protocol

TER - Translation Edit Rate

TM - Translation Memory

TT - Target Text

WM – Working Memory

Part I:

Research Context

Chapter One:

Introduction

1.1 Background

The research questions guiding the current study grew largely from exchanges in the Centre of Next Generation Localisation [1] between the researcher, colleagues at Dublin City University and industrial partners at the localisation department of the Symantec Corporation, Dublin. In these exchanges, controlled language (henceforth CL) re-emerged consistently as an issue of importance in translation, terminology management, data quality assurance, and machine translation (henceforth MT), both in research and development in academic contexts, and in current industrial practice. Throughout each phase of the project, there was close involvement of academic and industrials partners, where great contributions were made, especially in terms of resources, training, and feedback generously provided to the researcher, all of which helped to shape the current study.

Controlled languages (CL) have been around for many decades (e.g. Ogden 1930), yet it is only in recent years that they have received significant academic attention (e.g. Barthe 1998, Carl 2003, Geert *et al*. 2002). CL has been investigated, in particular, in the context of MT workflows (Roturier 2006) where the CL is used to constrain the lexicon and grammar used in source texts in order to facilitate automatic translation of those texts. The ultimate aim in using CL in these contexts is to improve MT output and, where applicable, reduce the amount of post-editing required (Allen 2003, O'Brien 2006).

Industrial application of CL has tended to outpace academic activity in the area, however, and although many translation/localisation vendors use CL under various names and guises, the associated body of research leaves many unanswered questions. Programmes such as the Controlled Language Application Workshop (e.g. Mitamura *et al*. 1998) series have gone some way towards addressing the lack of academic research in the area, but it remains the case that a great deal of CL and MT research takes place in industrial settings,

[1] The Centre for Next Generation Localisation is a research centre that brings together around one hundred researchers from academia and industry, all with an interest in the technologies used in translation, localisation, and personalisation. It is funded by Science Foundation Ireland. See http://www.cngl.ie.

and access to results from such research may be restricted for commercial reasons (see O'Brien 2003).

Studies of CL and MT can take very different approaches. Given its linguistic nature, CL has been an area of research in translation studies (e.g. Roturier 2006, Aranberri Montasterio 2009), while MT research tends to be the preserve of computational linguistics and computer science (Wilks 2009). A number of recent interdisciplinary studies have, however, shown how the two areas can complement each other (e.g. Sun 2010, Tatsumi 2010, Way and Gough 2005).

At the same time, researchers such as O'Brien (2006) have shown how methods used in human-oriented translation process studies (e.g. Alves 2003, Kenny and Opitz 2000) can enhance studies of the interaction of users with translation technologies such as translation memories and MT. With technological advances in key-logging (Jakobsen 2006, Leijten and Van Waes 2005), eye tracking (Dragsted and Hansen 2009, Jakobsen and Jensen 2008), and brain imaging (Andonova *et al*., 2009, Gerganov *et al*. 2008), the hitherto *black box* of the translator, or indeed post-editor, evaluator, or user, has become somewhat more accessible to researchers, who tend to use these technologies in conjunction with methodologies and ideas from domains such as cognitive psychology and psycholinguists.

The study reported on in this thesis marks a further contribution to an emerging interdisciplinary literature that combines insights and methods from the areas of CL, MT and translation process research. It is methodologically innovative in that it integrates methods more commonly associated with translation process research into the investigation of the reception of a particular type of translation product—MT output. In so doing, it draws on aspects of psycholinguistics and cognitive psychology, in both its theoretical foundations and experimental practice (Eysenck and Keane 2008).

More specifically, the current study examines the impact of CL on the readability and comprehensibility of MT output presented to human users of technical support documentation. Readability is operationalised as the extent to which a text can be easily read in terms of its formal linguistic elements. It is measured using Flesch and LIX readability indices, on the one hand, and human

judgements of readability on the other. Comprehensibility is defined as how easy a text is to understand and is measured using a combination of human judgements and a recall test, which allows comprehension to be tested indirectly. These textual and human evaluation methods are supplemented by data gathered by an eye tracker as participants read MT output on-screen. Certain eye tracking metrics are commonly used as indicators of cognitive effort, such as fixations and pupil dilation.

In using mixed methods and triangulating results from readability indices, human evaluations, recall tests, and eye-tracking studies, the researcher declares an intention to view the research question 'in the round'. The research thus aims to be as holistic as possible, while still responding to the need to operationalise individual concepts (readability, comprehensibility, etc.) in objectively valid and measurable ways. The use of such mixed methods is typical of research that draws on the tenets of philosophical pragmatism; research methods are chosen on the basis of how useful they are, given a particular set of research questions - rather than on the basis of a prior commitment to a particular theory or method.

Finally, the current research also prioritises ecological validity. Therefore it was vital that materials used in the current study were drawn from contemporary industrial use. Thanks to the cooperation of Symantec, such ecological validity was ensured. The use of data from this source also makes the findings of this research more relevant to such industrial parties.

1.2 Research Questions

The main research question of the study asks:

Does the implementation of linguistic pre-processing in the form of a controlled language rule set result in higher levels of readability and comprehensibility in Statistical Machine Translation output?

As already indicated, readability is (partly) operationalised in this research using indices familiar from the literature, namely Flesch and LIX. This study thus poses the following more specific question:

- *Does implementation of CL result in improved scores as measured by the traditional readability indices Flesch and LIX?*

Given that the research attempts to use eye tracking data as a source of information about cognitive processes involved in reading MT output, it also asks:

- *Are differences in eye tracking measures reported between the uncontrolled and controlled conditions?*

Readability is further gauged using a human evaluation, and comprehensibility is measured using a human evaluation and through a recall test. The study thus asks:

- *Do post-task human evaluation and recall testing show an improvement in readability and comprehensibility after implementation of CL?*

Another set of questions examines the relationships between the various metrics used in the research:

- *Do all of the above measures correlate and yield consistent findings?*
- *What is the relationship between human and machine evaluation of MT in this context?*

These research questions are further broken down in Chapter Three on Methodology.

1.3 Thesis Structure

This thesis is divided into four parts, the first of which, *Part I*, provides the context and rationale for the current study by means of an introduction and an outline of the proceeding chapters. It then moves to a review of relevant literature (Chapter Two) to provide a detailed description of research carried out in several disciplines, which informs the research questions and design of the current study.

Part II, the next two chapters (Three and Four), represents the two main phases in the methodological and chronological development of the study. Chapter Three discusses the methodology adopted in this study. It begins with an outline of the theoretical underpinnings of the methodology and goes on to operationalise important concepts. It also gives more concrete explanations of the methods employed in the study. Chapter Four discusses the application and testing of the methodology in a pilot study. It also provides the results of the pilot and identifies necessary refinements of the methodology in preparation for the main study.

Chapter Five represents the second phase of the study, in which methodological refinements are implemented, and the main study is prepared and carried out. It goes on to describe the statistical analyses that were applied to data elicited in the main study, and how the quality of eye tracking data was assured. It also describes how data and research instruments were prepared for use in the main study.

Part III (Chapter Six) provides a detailed account of the results of the main study under three headings (textual variables, eye tracking metrics, and human evaluation variables). It also discusses the correlations between different metrics, and concludes with a review of the research questions vis-à-vis each question's hypothesis and its result.

Part IV, the final part of the thesis (Chapter Seven) sums up the study's main findings, discusses its strengths and weaknesses, and the contribution it makes to scholarship in the relevant domains. It presents the implications of the findings for those who wish to implement CL in industrial scenarios, and suggests fruitful avenues for future research.

Chapter Two:

Literature Review

2.1 Chapter Overview

This chapter reviews literature relevant to both the pilot and main studies, and additional necessary background information. It begins with an exploration of the concept of readability (section 2.2), highlighting the readability debate, focusing on the two indices used in the current study, and describing the link between readability and comprehension. Section 2.3 presents the concept of controlled language, and explores commonalities between several controlled languages and reviews associated studies. Section 2.4 introduces machine translation, outlining how machine translation systems have developed, and presenting the broad categorisation of MT systems into rule-based and statistical machine translation systems. It describes how controlled language has been implemented in MT workflows and closes with a review of practices in machine translation evaluation and automatic evaluation metrics. Section 2.5 reviews the eye tracking literature and especially those sources that focus on translation process studies. Lastly, section 2.6 explores the cognitive aspects of reading, translation, and comprehension, beginning with descriptions of human memory systems, and going on to discuss memory decay and recall, and the use of Think-Aloud Protocols as a research methodology.

Each section ends with a summary of the main points, and an overall chapter summary closes the chapter.

2.2 Readability and Comprehensibility

2.2.1 Section Overview

This section is concerned with research into readability and comprehensibility. It first charts the development of readability research, and then moves on to discuss the strengths and weaknesses of readability indices. The Flesch and LIX indices are described in detail as they are the two metrics of readability employed in the current study. The link between readability and comprehensibility is then examined, and a more expansive review of works specific to comprehension is provided. A section summary provides a synopsis of the main points concerning readability and comprehensibility.

2.2.2 The Development of Readability

The concept of readability has existed for some time. However, it is only in the last 80 years that it has been thoroughly researched. Much of the groundwork was carried out in the USA in the first half of the 20th century, paving the way for both progress and contention. The main goal of readability measurements is to provide an accurate indication of the difficulty of a text; this concept itself is, however, subject to debate.

Klare (1974) distinguished between two approaches to readability: measurement or prediction; measuring involves actual readers, predicting uses formulae. He defines a readability formula as a formula that "uses counts of language variables in a piece of writing in order to provide an index of probable difficulty for readers" (ibid., p. 64). Early readability research was carried out by Lively and Pressey (1923) who measured difficulty by assigning the Thorndike frequency[2] number to each different word in a given text and finding the average of those numbers to come to a final measure of readability. Texts with a lower number were more difficult than those with higher numbers. This test focuses on vocabulary difficulty as a factor in readability.

Vogel and Washburne (1928) proposed the Winnetka formula, which correlated elements of text difficulty to specific reading levels. They used four elements for defining difficulty: number of different words present in the text, total number of words in the text, total number of prepositions, and number of simple sentences. In this study Thorndike's list was also used, as were sample sentence sets.

Gray and Leary (1935) investigated readability in their experiment concerning the average comprehension of tests by a group of 800 adult readers. Their results helped to develop a formula using five elements: number of different difficult words, number of pronouns, percentage of different words, average sentence length, and number of prepositional phrases. However, Lorge (1939) examined Gray and Leary's formula and came to the conclusion that only

[2] The Thorndike list (Thorndike 1921) was one of the first extensive word frequency lists in English and provided one of the first objective standards of word difficulty.

three of the elements were valid: average sentence length, number of difficult words, and the number of prepositional phrases.

Flesch (1943) found that both Lorge's, and Gray and Leary's formulae were not appropriate when used with adults with more than a limited reading ability. He constructed a formula using three elements: average sentence length, the number of affixed morphemes, and the number of personal references. He later defines readability as "comprehension difficulty" (Flesch 1948) and proposes a revised formula based on three language elements: average sentence length in words, number of affixes, and number of references to people. He demonstrates the accurate and widespread use of his early formula but develops its replacement. The new formula takes the factor of human interest into account on the assumption that word and sentence length directly influences respective complexity. This revised formula is used in the current study and will be described in greater detail in the next chapter.

Dale and Chall (1948) highlighted two shortcomings of the first Flesch formula for readability, namely its counting of affixes and personal references. To solve these problems the authors hypothesise that: a larger word list would be more accurate than affixes, personal references are not important to readability, and a more efficient formula could be developed from a "word factor" and a "factor of sentence structure" (ibid., p. 15). The proposed formula was tested with human informants with 376 texts and Dale's own 3,000-word list. Results showed that average sentence length is an important factor in assessing reading difficulty.

Gunning (1952) proposed the Fog formula, which is similar to Flesch's approach. In the former case, however, the percentage of polysyllabic words (i.e. words with more than three syllables) is taken into account. Gunning's formula was later modified by Kwolek (1973) who used it to measure readability based on sentence length and the percentage of hard words in the text, where words with more than three syllables, symbols and abbreviations are classed as 'hard' words.

Klare *et al*. (1955) conducted a study to examine the effect of prior knowledge and other variables on retention and acceptability of technical texts. They found that more readable texts resulted in higher levels of retention,

increased number of words read in a given time, and greater reader acceptance. Additionally, it was found that "style difficulty appears to affect immediate retention of subjects who are naïve regarding material, subjects who have considerable knowledge of the material may profit little if any from an easier style of material" (ibid., p. 294). Similar experiments have also been carried out by Entin and Klare (1985).

Bormuth (1966) used cloze tests to determine the difficulty of 20 texts read by the 675 people who participated in this study. Cloze testing provided a new means to validate readability formulae and also brought comprehension into focus (see below). This type of testing became more popular and has been used in its original form and modified in some cases. In the original form of cloze testing, subjects were given a sentence containing a blank space which required a missing word to be filled in; subjects would either know the correct word by the context or not. The results of the above study show that readability formulae can accurately predict difficulty across various levels of reader ability and can successfully predict difficulties in individual words, clauses and sentences. Additionally, Bormuth (ibid.) found that nonlinear techniques are required in readability formulae and that further development of linguistic variables would yield great improvements in readability formulae.

McLaughlin (1966) showed that word and sentence length are the most accurate linguistic measures to predict readability. He also states that "in English, word length is associated with precise vocabulary" (1969, p. 640), whereby long sentences require more 'immediate memory' to allow the reader to construct the meaning of the entire sentence, and therefore require more cognitive effort. McLaughlin (ibid.) counts polysyllabic words and uses this factor in his formula, a formula that he argues is faster and easier to use than others and also more accurate.

Fundamentally, the majority of readability measures have focused on two main factors: the familiarity of the semantic units (words or phrases) used and the syntactic complexity of the sentence structure. For example, syntactic complexity is usually measured by sentence length and/or the number of clauses or phrases present (Drum *et al*. 1981, Klare 1984). Drum *et al*. (1981) state that the number of clauses per sentence is a more accurate measure of syntactic

complexity than sentence length. Moreover, vocabulary is usually measured by counting the number of syllables or letters in a word, or the word's location on a frequency list such as Thorndike's list. However, such lists present problems as they are language- and domain-specific.

In addition to this, various general guidelines have been proposed on how to produce readable texts without having to use indices (or, if an index is used, the guidelines should theoretically improve the text's score). DuBay (2004, p. 2), for example, states several rules of readability, some of which, it is argued here, are vague and ambiguous themselves:

- Use short, simple, familiar words;
- Avoid jargon, use culture-and gender-neutral language;
- Use correct grammar, punctuation and spelling;
- Use simple sentences, active voice, and present tense;
- Begin instructions in the imperative mode by starting sentences with an action verb;
- Use simple graphic elements such as bulleted lists and numbered steps to make information visually accessible.

It is interesting to note that although most of the research in this area has been carried out with English texts, several other indices exist for other languages; French being of particular interest to the current study. Kandel and Moles (1958) adapt the Flesch Reading Ease formula for use with French and de Landsheere (1973) creates a French version of the same formula. Henry (1973) also proposes a formula that could be used manually or by computer. A fuller description of the Kandel and Moles' adaption will be given in the next section as it, and the LIX formula (see section 2.2.4), will be used in the current study.

With respect to text domain, most research has been carried out using general, i.e. non-domain specific text. However, some readability indices have been applied to specific domains, for example, accounting (e.g. Smith and Taffler 1992), which, as already indicated, presents challenges in terms of the employment of word frequency lists and domain knowledge.

In attempting to develop an appropriate means of measuring readability, other issues arise such as the reader's involvement with the text, motivation, linguistic ability, world and domain-knowledge, all of which are obviously of great significance to the reader's understanding of the text and are noted in the literature (Shnayer 1969, Schriver 1989, Carrell 1987).

From a practical point of view, the formal elements of a text can be controlled e.g. with a controlled language or style guide, whereas it may not be possible for an author to control who reads the text and the way in which way it is read. It could be contended that certain readers would have an interest in certain texts. For example, some reading may be task-oriented, such as when someone reads a manual to fix a printer. And domain-specific text can become more familiar to the general population via media and world events, e.g. financial and Internet-related vocabulary.

Given the multiplicity of possible approaches to 'readability' it is advisable at this point to define the concepts of readability and comprehensibility as they will be understood in the current study. Readability is defined here as the extent to which a text can be easily read in terms of linguistic elements (such as number of syllables, number of words and sentences), i.e. it is operationalised as a text-dependent attribute. Comprehensibility (discussed in section 2.2.5) is defined here as the extent to which a text is easy to understand. It is classified here as an attribute of the text which is relative to and dependent upon the reader, i.e. it can change depending on the reader (reader-dependent) whereas readability is anchored to the text. The process of reading, therefore, involves a combination of both factors (plus additional factors such as motivation, time, etc. which are also discussed below). Further detail on this operationalisation is given in the next section (2.2.3), and in Chapter Three.

2.2.3 The Readability Debate

Criticism of readability studies has been documented for as long as such research has been carried out (e.g. Hargis 2000). Davison and Kantor (1982) criticise readability measures and investigate their accuracy in their own study. They find that editing a text to aim for a certain readability grade, e.g. reducing sentence length, has consequences for the way in which information is given and interpreted. They argue that topic, focus, inference, and point of view are also of importance.

Duffy (1985) criticises established indices, his main criticism being that the variables on which the indices are built are not the most accurate means of measurement, for example, sentence or word length. Schriver (1989) highlights the need to take other factors into account e.g. reading skill, subject knowledge, motivation, context and purpose of reading. The passages used in the comparison with the text to be tested also present a problem. In many cases these passages were very short and therefore do not provide a comprehensive baseline measure (see Klare 1984). Additionally, the criterion passage used in indices has varied, thus presenting possible issues in consistency.

Homan *et al*. (1994) find differences in their subjects' responses to items estimated to be equal to their readability level and those estimated to be above this level (ibid., p. 356). They also stress that text creation be employed in conjunction with readability indicators by making use of an average readability level for the text as a whole, thus allowing for individual items to be above or below the intended level (ibid., p. 349).

Redish (2000) highlights the weakness of readability indices and promotes the case of usability studies. She draws attention to the fact that most indices were designed for American grade-school students and are therefore not appropriate for use with adult readers, and that most of the indices and associated research is out of date (ibid., p. 133). She also states that indices usually count certain text features, such as sentence length, because they are easy to count (ibid.). She draws attention to features that are not counted, for example: suitable content for the audience, text organisation (headings, index, tables of contents, layout, and familiarity with vocabulary).

Schriver (2000) reports on criticism of readability indices from the 1970s and comments on their validity today. She criticises indices that take syllables per word and words per sentence into account, thereby penalising texts that do not make extensive use of full stops – thus simply adding full stops where other punctuation is more appropriate is said to increase readability. She also describes cases where writers "write to the formulas" (ibid., p. 140) and impose limits on word length.[3] This usage had already been mentioned by Klare (1984), who pointed out that the indices should not be used as a guide for text production in the first place.

Giles and Still (2005) highlight the problems that the more commonly used readability indices pose to technical writing. They propose the Golub Syntactic Density Formula, which examines sentence syntax at a deeper level than the methods of syllables per word and words per sentence, i.e. at the clausal and phrasal level. However, such a method shares some of the aforementioned shortcomings of readability indices in that it relies wholly on linguistic information.

In defence of the indices, Klare (2000) provides further information and advice on the correct use of readability indices. In addition, he points out that although some were intended for student readers, they may be applied to adults also. He (ibid.) acknowledges the drawbacks of several indices and the criticism they received and moves onward to focus on producing readable documentation for the domain of computing, where controlled language is mentioned as a being of value.

Collins-Thompson and Callan (2005) provide an example of readability measures being adapted to suit more modern needs. They attempt to develop a way for information retrieval systems to match texts to the reading ability of student users. They find that traditional readability measures, such as Flesch-Kincaid[4] are not applicable to Web pages, and adopt a statistical model to classify Web pages according to reading difficulty. They find that an approach based on

[3] There is an interesting parallel here with controlled language, which will be discussed later.

[4] The Flesch-Kincaid formula converts the Flesh score into grades based on the American education system.

vocabulary analysis is accurate in this case, yet such use of vocabulary presents issues of domain-specificity.

Connatser (1999) argues that the use of traditional readability indices such as Flesch is redundant in the domain of technical support documentation as "most audiences of technical documents read to do" (ibid., p. 284). He concludes that the implementation of usability testing is therefore more appropriate in this context; other researchers such as Hargis (2000) come to a similar conclusion.

Overall, it appears that evidence exists for the accurate use of readability indices to measure linguistic phenomena. However, the interaction with the reader presents a confounding variable, yet the text cannot be read without a reader in the first place. Other research points to the inadequacies of readability indices for some purposes, and the employment of additional methods such as usability testing to supplement, or even replace, readability tests, has been widely suggested. Therefore, it is not advisable to rely solely on readability indices for concrete results, unless correlations have been found with other measures beforehand. If such correlations are found then readability indices can be reused in similar conditions without additional methods. On the other hand, as readability indices are extremely resource-cheap, it would not be advisable to ignore them completely, once the user is sure of what they do and do not measure.

In conclusion, for the purposes of this research, and in line with most uses of the term 'readability', it is understood here as something that can be largely controlled a priori, by means, for example, of a controlled language or style guide. Readability can also be measured a posteriori using an index. While such an approach to readability may appear to be reductionist, it has the merit of being highly operationalisable, once suitable features of texts have been isolated that can reasonably be assumed (on the basis of prior research) to correlate well with the ease with which humans (with certain attributes) can understand a text. Given this understanding of readability we can say that readability indices measure rather than predict readability, which is an attribute of texts only (Klare 1974-5). To the extent that they can predict anything, it can be stated that they are a factor in predicting comprehensibility (see below), which depends crucially on attributes of the reader as well as the text; but a high (i.e. good) readability

index by itself is neither necessary nor sufficient to ensure that a given human will easily understand a text. Such a position allows the metrics which are designed to solely measure textual elements to do so. The aforementioned additional human factors are not easily quantifiable and will be addressed elsewhere, in particular in section 2.5 on eye tracking.

2.2.4 The Flesch and LIX Indices

The application of readability measures to French is rather more recent when compared to research on the English language. As already indicated, Kandel and Moles (1958) adapted Flesch's formula for use with French text. Later, de Landsheere (1973) added to this adaptation and there are currently three indices that have been validated for use with French (Henry 1973).

With further regard to readability specific to French, Richaudeau and Staats (1981) present four essential principles, which echo Klare's pillars and state that readability increased when the proportion of "everyday, short, concrete, personal words" increases, "average sentence length decreases, sentence structure becomes more simple" and "whenever a subject reading a sentence altered by the cloze procedure can guess a larger proportion of missing words" (ibid., p. 503). Through their work they suggest that neither sentence length nor closeness to kernel structures is an essential factor of readability. As seen in the literature, the most widely used metric for French text is the adaptation of the Flesch formula by Kandel and Moles (1958), which will be used in this study for this reason and also due to the ability to compare it to the Flesch scores obtained from the English version of the translated text (discussed in Chapter Three).

Björnsson (1968) developed the LIX formula in an attempt to measure readability across languages. This formula focuses on word and sentence factors such as length and the score obtained is then compared to a scale indicating difficulty. The formula was tested with English, French, German, Greek and Swedish texts. The usefulness of this formula is demonstrated by several researchers such as Lewis *et al*. (1986). The LIX formula is of particular interest to the present research as it deals with readability in more than one language, in this case, English and French. Formal definitions of both Flesch and LIX will be presented in Chapter Three.

2.2.5 The Link between Readability and Comprehensibility

Throughout the literature, the topics of readability and comprehensibility are intertwined and comprehensibility is often not discussed explicitly as sometimes it can be the case that readability assumes comprehension. Yet many researchers distinguish clearly between the two concepts. Klare (1974) distinguishes between readability and comprehensibility by stating that a readability formula is a "predictive device", whereas tests to measure comprehension are not. He also describes the move to using cloze testing which can "yield higher predictive validity coefficients" (ibid., p. 66) between readability and comprehensibility and has been used in other studies (e.g. Taylor 1953, Bormuth 1969)

Klare (1976) later examined 36 studies that attempted to improve text comprehension by increasing the text's readability score. He reports that about half of these attempts were successful and but still had to incorporate substantial changes to improve their score, i.e. corresponding to an average of 6.5 grade levels (based on the American education system of grade levels). Charrow and Charrow (1979) carried out a similar study of legal documentation and found that when they revised texts to increase comprehension, which consequently showed an increase, readability scores fell. Such a finding highlighted the complex relationship between readability and comprehensibility.

Harrison (1980, p. 33) makes a clear distinction between the two stating that readability is a characteristic of the text whereas comprehension is one of the reader, whereas Adelberg and Razek (1984) see no difference in a text being readable and it being understandable, while Jones (1988) states that comprehensibility is reflected in readability. Smith and Taffler (1992, p. 85) state that "comprehensibility can be different to readability and the latter might frequently be used erroneously as a proxy for the former". They find that readability and comprehensibility (called understandability by the authors) are different concepts in an experiment that used readability indices in conjunction with cloze testing and suggest that "understandability is related both to complexity of context and to education and experience" (ibid., p. 93). Although the target audience and domain of this experiment is specific, the methodology

used provides an interesting insight as it deals with both readability and comprehensibility. Lastly, they state that comprehensibility and readability cannot be measured by the same index due to their inherent differences and that comprehensibility concerns the complexity of the text content and the education and experience of the reader.

However, Chall and Dale (1995) argue that most of the features of a given factor such as readability and comprehensibility are highly correlated with each other, so one estimate for each factor is sufficient. Support for this can be found in numerous studies, where vocabulary difficulty and average sentence length are the two features that have been found to be the most consistently and strongly associated with comprehensibility (Chall 1958, Klare 1963).

From the above research the somewhat unstable line of progress with respect to developing an accurate readability measure is evident. For the most part, an accurate measure cannot be agreed upon and the shortcomings of any formula based on linguistic criteria can clearly be seen. Furthermore, the concept of comprehensibility has from time to time been blurred and overlapped with that of readability, resulting in a lack of a clear and concise definition to this day where it seems that both concepts and their use remain rather subjective and inconsistent. Given the differences between formal elements of a text, on the one hand, and other factors that impinge upon a reader's ability to understand a text (motivation, domain knowledge, etc), on the other, it is not surprising that a single concept cannot easily subsume both, and it is even less surprising that a single metric or index cannot capture both concepts.

2.2.6 Section Summary

This section focused on the topics of readability and comprehensibility. It presented the development of readability research over recent decades, and then discussed the strengths and weaknesses of the measurement of readability by means of indices. It has been shown that although there are several drawbacks to readability indices, when used in indicated circumstances, and especially in conjunction with other methods, these indices can provide useful information about texts and how they are likely to be perceived by readers of different ages and abilities. The two indices used in the current study, the Flesch and LIX indices, were then described in further detail. Lastly, descriptions of comprehensibility grew from the discussion of readability and additional research in this area was reviewed. It was evident that although the boundaries between readability and comprehensibility can appear to be somewhat blurred, consensus can be reached around fundamental aspects of these concepts, which laid the way for the operationalisation of the concepts in the current study.

2.3 Controlled Language

2.3.1 Section Overview

This section is concerned with the topic of controlled language. Firstly, it defines what is meant by a controlled language in the context of this study, and describes identifiable commonalities of controlled languages. Secondly, it reviews research on controlled language and closely related topics. Lastly, a section summary provides a brief recap and acts as a bridge to a discussion of machine translation.

2.3.2 Controlled Language Commonalities

A controlled language can be defined as "an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar, and style" (Huijsen 1998, p. 2). The application of such constraints to a text aims to: improve comprehensibility, ease of processing and post-editing (of the machine translated text), and ensure consistency and quality (Douglas and Hurst 1996). In relation to translation, the use of controlled language input has been shown to improve the quality of the output, whether the translation is done by humans or machines (Nyberg *et al. 2*003).

Controlled languages can be divided into two types: those that aim to improve the ease with which human readers can understand the text (human-oriented controlled languages or HOCLs); and those that attempt to increase translatability, i.e. the ease with which a text is translated (see Gdaniec 1994, Bernth and Gdaniec 2001), and comprehensibility of a text by natural language applications, not just machine translation systems (machine-orientated controlled languages or MOCLs).

Both types of CL share common objectives, namely to reduce ambiguity and increase readability/translatability. Ambiguities can be, for example, lexical, structural, referential, and syntactic (Hutchins 2003). Huijsen (1998, p. 2) differentiates between the two, stating that "writing rules for the machine-orientated controlled languages must be precise and computationally tractable", e.g. "do not use sentences of more than 20 words", whereas HOCL rules may be vaguer, for example, "make your instructions as specific as possible, and present new and complex information slowly". Similarly, Clémencin (1996, p. 32) states that MOCLs attempt to "simplify and normalize the linguistic content of documents in order to match the capacities of automatic translation tools", whereas human-orientated CLs are not adequate for Natural Language Processing (NLP) due to their lack of formalisation and explicitness (Lux and Dauphin 1996, p. 194). A further distinction between MOCLs and MT oriented CLs (MTOCLs) is observed by Vassiliou *et al*. (2003), who explain that an MTOCL can be optimised for use with a particular MT system, which is the case for the CL rule set used in the current study (discussed in Chapter Three).

O'Brien (2003) provides the only published review of eight CLs and classifies the rules by their primary function as follows:

- Lexical;
- Syntactic;
- Textual - subdivided into 'text structure' and 'pragmatic'.

Her study found that only one rule was shared by all the CLs she reviewed and this was the rule promoting short sentences - illustrating the unique nature of the rules contained in the CL rule sets examined. O'Brien suggests that this lack of overlap is due to differences between the objectives of the CLs, and the MT systems and language directions in use, and to the influence of corporate writing rules/authors, and general subjectivity (ibid., p. 111).

Finally, the concept of sublanguage is worthy of mention here to avoid confusion between it and CL and to supplement research mentioned in the previous section. A sublanguage is a type of language that has developed naturally in a specific domain, and uses particular vocabulary and grammar, that may or may not be used in the language in general. Roturier (2006) states that the difference between a sublanguage and CL is that a sublanguage is not artificially controlled or created "it just happens to have a limited number of linguistic features" (ibid., p. 47). An interesting example of sublanguage is found in the language of weather forecasting, as capitalised upon by the TAUM-METEO project in which an MT system was developed to translate weather forecasts between English and French (Isabelle, 1987, Lehrbrger and Bourbeau 1988). The main shortcoming of sublanguages according to Nirenburg (1987) is that it is difficult to find a completely self-sufficient sublanguage that would be useful to a given domain or purpose. Therefore, it is typically necessary to artificially create a CL rule set as a subset of language for a particular use/domain.

2.3.3 Reviewing Controlled Language Research

Research concerning controlled language began as early as 1930 with Ogden's Basic English (1930), which comprised 850 words and was designed as an international language and English learning tool. Further research continued over the following decades, the most relevant of which will now be described in detail. Although English has been the focus of this research, examples of CLs in other languages exist: GIFAS Rationalised French (Barthe *et al.* 1999), ScaniaSwedish (Almqvist and Sagvall Hein 1996), and Siemens-Dokumentationdeutsch (Schachtl 1996).

Nyberg *et al*. (2003, p. 261) describe Caterpillar Fundamental English (CFE) as an example of a HOCL. It is "intended for use by non-English speakers, who would be able to read service manuals written in CFE after some basic training". An example of an MOCL can be seen in Fuchs and Schwitter (1996) and Kaljurand (2008). The latter describes Attempto Controlled English (ACE) as "a subset of English, such that each sentence in the chosen subset is interpreted unambiguously, relating the sentence to a unique form" (ibid., p. 1). ACE is intended to be expressive and simple enough to be easily used, while remaining a natural subset of English. ACE is an example of a CL that is designed to improve the comprehensibility of a text by programs using logic programming or artificial intelligence components (Fuchs and Schwitter 1996) and consequently has a low number of permitted structures. These aspects of ACE address an important factor in the success of a CL: it must be rigid enough to fulfil its objective, but also sufficiently easy to use and not too restrictive in the expression it allows its writers. Fuchs and Schwitter (1996, p. 3) state that "on the one hand this subset should be expressive enough to allow natural usage by domain specialists, and on the other the language should be accurately and efficiently processable by a computer".

Adriaens (1994, p. 79) describes the SECC (Simplified English Checker/Corrector) project, where Simplified English is defined as "a subset of regular English, consisting of Alcatel Bell's COLEX (a restricted regular English vocabulary), COTECH (a restricted technical English vocabulary from the domain of telephony) and COGRAM (a restricted grammar)". It consists of a core

vocabulary of 1,500 words, a set of writing rules for grammar and style, and words chosen for their simplicity and commonality with other European languages. The application of CL has been especially researched in relation to technical texts (see Huijsen 1998, Knops and Depoortere 1998, Means and Godden 1996). An example of this is described by Spyridakis *et al*. (1997) in that it is designed to ensure greater readability and consistency with a focus on technical documents. They (Spyridakis *et al*. 1997) investigate SE's impact on translatability and compare SE versus non-SE texts being translated from English into Spanish and English into Chinese by native speakers of the target language who are all novice translators. Using accuracy, style, comprehension, mistranslations, and omissions as metrics, they found that subjects who translated SE produced higher quality and scored higher on most of these measures than subjects who translated non-SE texts. They claim that CL produces varying results with different language pairs and suggest that the benefits for English-Spanish were clearer than was the case for English-Chinese, because Spanish is linguistically more similar to English.

Mitamura and Nyberg (1995) describe KANT (Knowledge-based, Accurate, Natural-language Translation) Controlled English as a CL with constraints on: lexicon, complexity of sentences (to limit parsing during source analysis), and the usage of a mark-up language that supports the definition of domain-specific terminology without increasing ambiguity in the text. Their experiments with this CL and the KANT MT system showed improvements in the quality of the translation.

AECMA's (European Association of Aerospace Industries) Simplified English comprises of a restricted vocabulary of 1,565 words and 57 rules for usage (Unwalla 2004), e.g. – sentence length limit 20 words or 25 for descriptive text, paragraph limit of 6 sentences, compound noun length limit of 3 words. It was designed to allow for the inclusion of a level of technical information needed for aircraft support and maintenance on an international level. Hoard *et al.* (1992) illustrates AECMA SE[5] in use and states that although it is difficult to implement and maintain and use, it has proven to be a successful application of

[5] After the merger of AECMA with two other associations to form ASD, the name of the CL became ASD Simplified Technical English, Specification ASD-STE100.

CL to a specific domain of industry. The implementation of the CL was aided by Boeing's Simplified English Checker, a program that uses 350 English grammar and parsing rules to help the writer of a text to adhere to the CL rules and guidelines and suggest alternatives where appropriate.

Unfortunately, due to confidentiality reasons, some results of studies relating to CL were not published fully; some of the CL rules used also suffered the same fate and have never been made completely public, e.g. those of Alcatel Bell, Caterpillar, General Motors, Sun Microsystems, and Xerox (Bernth and Gdaniec 2001).

While the focus in most research has been on the various CL rules and their adequacy for their intended purpose, few researchers have taken the impact of individual rules into consideration, probably due to the difficulty of such a task and unavailability of required materials due to proprietary usage. Nyberg *et al*. (2003, p. 257) state that "it is unclear what the contribution of each individual writing rule is to the overall results of the CL". O'Brien and Roturier (2007, p. 1) find "rules governing misspelling, incorrect punctuation, sentences longer than 25 words, and the use of personal pronouns with no antecedent in a sentence" were most effective in the context of improving comprehensibility of post-edited MT output. This is echoed by Roturier (2006), who also identified the most effective rules as consistent spelling, the avoidance of unusual punctuation, and a restriction of sentence length to 25 words.

In addition to the development of the CL itself, programs to help the writer adhere to the CL have also been created. Examples are Acrocheck (www.acrolinx.com) and Eurocastle (Clémencin 1996). Other CL checking environments have been used as an alternative to the more popular word processing programs. Power *et al*. (2003), for example, present a system that can produce multiple expressions of the same input in multiple languages, so the author can choose alternative expressions to satisfy the CL, obviating the need for correction.

Bernth (1997) describes EasyEnglish, a similar tool that highlights ambiguity and complexity to the writer. As with other checkers, both grammar and spelling are examined and suggestions are made by the application as to how mistakes can be rectified. EasyEnglish also uses a Clarity Index by which a

text is rated and only certain scores are accepted for publication. Such approaches of using tools to help authors adhere to a set of given rules and guidelines have played an important part in the success of CL. Reusability of entire files has also improved due to the adoption of CL rules sets, e.g. at Caterpillar (Hayes *et al.* 1996) and General Motors (Means and Godden 1996). From a terminology point of view, the use of a CL ensures consistency and proves to be both cost and time saving (Allen 1999). Furthermore, when used in conjunction with other tools, a CL can be of even more use.

Unfortunately, there are certain drawbacks associated with the use of CL. Govyaerts (1996, p. 139) notes that deploying a large set of CL rules is sometimes difficult due to time and resource constraints, even with the use of a CL checker. The rules imposed by a CL can reduce or force expression, make the writing task overly complex, and resistance to or lack of familiarity with the rules can cause difficulties with writers. Van der Eijk *et al.* (1996, p. 64) notes how "grammar restrictions often can only be expressed in a linguistic jargon that is not always easy to explain to authors, who normally are domain experts with no or limited linguistic background". In addition, the objective and adequacy of each rule should be taken into account; Huijsen (1998, p. 12) states that "some writing rules may even do more harm than good". It is evident that controlling language in certain contexts can be fruitful; however, the above problems highlight the need to find middle ground in order to successfully use a CL in any scenario.

Reuther (2003, p. 131) has argued that a link exists between translatability, i.e. the quality and ease of translation (Spyridakis 1997), one of controlled language's main aims, and readability and comprehensibility. She finds that readability rules are a subset of translatability rules and "translatability ensures readability", whereas the reverse is "only true to some extent" (ibid., p. 7), although further investigation into this connection is necessary. There is on-going debate regarding the compatibility of these two goals. Bernth and Gdaniec (2001) showed that after applying MT-oriented CL rules their text corpus improved in clarity and translatability but reduced readability, a finding that is inconsistent with Reuter (ibid.) Reuter (ibid.) found that rules dealing with the lexicon and ambiguity proved the most important for improving readability, whereas rules dealing with ellipsis and typography had

little impact. In contrast, translatability relied mostly on rules dealing with ambiguity and ellipsis, whereas it did not depend on lexical rules.

A more recent study, mainly focusing on rules addressing readability but also including MT-oriented rules suggested that controlled texts were "easier to read, are viewed more favourably, and encourage better retention of keywords" (Cadwell, 2008, p. 50). A follow-up study, on the contrary, limited these results by showing that CL rules might be beneficial in terms of readability and acceptability for complex texts but not for easy texts (O'Brien 2009). More comprehensive studies on the relation between readability and translatability are required to shed light on these contradictory findings.

In attempting to address the question of CL's effect on translation, De Preux (2005) used error-severity scores. The results suggested that although the number of errors did not decrease with the implementation of CL, their severity was reduced. In another study, a significant improvement on the output of a commercial MT system using CL was reported (Roturier 2004). Output was classified as excellent, good, medium and poor. Excellent output is defined as being ready for review. In Roturier's context, poor output is discarded and the source sent to translators for traditional human translation. Good and medium quality output is sent to post-editors. It was found that excellent output doubled for all languages and medium quality examples decreased considerably when CL was implemented.

2.2.4 Section Summary

This section dealt with controlled languages. The commonalties of several controlled languages were described and relevant studies of controlled languages were reviewed. It is evident from the review of the literature that although CLs can vary greatly depending on the organisation's needs, e.g. technical writing or MT, there are identifiable similarities such as restricted sentence length. The proprietary nature of CLs and the lack of published research in the area were also highlighted as restrictions to the exploration of CL in the context of the current study. The CL used in the current study will be described in greater detail in the following chapter. As CL is often used in conjunction with MT, the latter will be the topic of the next section.

2.4 Machine Translation

2.4.1 Section Overview

This section focuses on machine translation systems. It provides a brief description of both rule-based MT and statistical MT. It goes on to outline a number of studies into the use of controlled language in machine translation workflows. The section ends with a discussion of the evaluation of machine translation systems by means of automatic evaluation metrics.

2.4.2 Machine Translation Systems

Within computational linguistics and computer science, machine translation denotes the use of computers to automatically translate text or speech from one natural language into another. Hutchins (2005, p. 1) highlights the growing need for MT stating that "there is just too much that needs to be translated" and "human translators cannot cope" with this ever increasing volume. From an industrial point of view, companies are constantly seeking ways to reduce translation costs, and MT as well as tools such as Translation Memory systems, can provide savings in terms of resources, and possible improvements in terms of consistency and reuse of content etc.

The early, and in retrospect, overly ambitious aim of MT was to achieve Fully Automated High Quality Machine Translation (FAHQMT). Today, MT researchers have more realistic aims and Somers (1997, p. 116) notes a shift of focus from aiming to attain FAHQMT to the "sudden interest in using MT to get rough translations". Problems faced by MT systems at the time remain today however, for example, ambiguity of language (Arnold 2003), and intrinsic features that are unique to a language and present difficulties to an MT system in its interpretation and translation of a given phrase (Forcada 2010).

MT systems can typically be divided into two main approaches: rule-based and corpus-based (also called data-driven). Rule-based systems (RBMT) use grammatical and lexical rules (often manually written) to govern the translation process, while corpus-based systems, such as statistical MT systems (SMT), are constructed based on large monolingual and bilingual parallel corpora.

Somers identifies a new emerging trend in the early 1990s "characterized by the preference of data-orientated models of language, derived by statistical or analogical methods, as opposed to rule-based models derived from linguistics" (ibid., p. 115). He comments on the appearance of hybrid systems, which are comprised of:

- Rule-based systems where the rules are derived more or less automatically from data;

- Analogy-based systems where the examples are generalised so as to take on the form of rules;
- True hybrid systems, where the alternative approaches work alongside each other.

(ibid., p. 117)

The focus of the present research is on a hybridisation of RBMT and SMT, discussed in more detail below, while the specific implementation of the MT system used in the current research will be provided in Chapter Three on Methodology. For a discussion of analogy-based, i.e. example-based, systems see Carl and Way (2003).

2.4.3 Comparisons of RBMT and SMT

There are two sub-types of RBMT systems: transfer-based, and interlingua systems. Although both systems work under the same concept of using an intermediate representation to encompass the meaning of the ST and render it accurately and fluently in the TT, transfer-based systems create an intermediate representation that is dependent on the language pair involved in the translation. Therefore the representation created by a transfer system for the English-French language pair would differ from that of the English-German language pair. Interlingua-based systems create a representation independent of the language pair and can therefore be said to be language-independent. Typically, the following stages are used in the translation process of RBMT (Hutchins and Somers 1992, p. 75):

- *Analysis* of the ST to extract linguistic information from the input (parts-of-speech, syntax etc.) via parsing;
- Creation of appropriate equivalent in the TT or *transferring*;
- Rendering of the text in the TL or *generation* (ensuring correct use of the TL).

As noted by Forcada (2010), most rule-based approaches to MT are transfer-based systems. Notable shortcomings of the RBMT approach are the scarcity of high quality bilingual dictionaries and grammars and the fact that creating new dictionaries/grammars can be expensive and time-consuming. The process of RBMT may not be wholly automated from run-time in that linguistic information may need to be input manually. RBMT systems tend to have difficulties with lexical ambiguity such as idiomatic expressions.

Within the corpus-based (data-driven) approach to MT, there are two main paradigms: statistical (Brown *et al.* 1988, Koehn 2009), and example-based (Nagao 1984, Carl and Way 2003). Here the focus moves to SMT systems, the currently dominant paradigm of MT research with an increasing presence in

commercial applications (Forcada 2010). The increase in interest in SMT can be attributed to the development of the algorithms for training the models used to calculate probabilities in the SMT approach. Aue *et al.* (2004, p. 1) note how "much research of late has been devoted to the invention and implementation of SMT systems". Others, such as Déchelotte *et al*. (2007, p. 1) state how SMT is the current "preferred approach of many industrial and academic research laboratories, each of them developing their own set of tools".

SMT centres on the principle that a probability can be assigned to any sentence in the SL being translated as a particular sentence in the TL. Figure 2.1 illustrates a typical SMT approach, where a series of algorithms are used to build:

- A target language model, learned from a monolingual corpus in the TL, and which assigns a probability to sequences of words (n-grams) in the target language, and
- A translation model, learned from a parallel corpus, and which assigns probabilities to translations for given SL n-grams.

A 'decoder' then finds the best possible translation pair, or n-best, from the proposed table, called the phrase table. The phrase table is populated by a list of all possible translation pairs in the given language pair; pairs that appear more frequently in the corpora are given higher rankings. A notable feature of pure SMT systems is that no linguistic information is used in the process or in the creation of the models; rather they are derived from a series of statistical techniques (Forcada 2010).

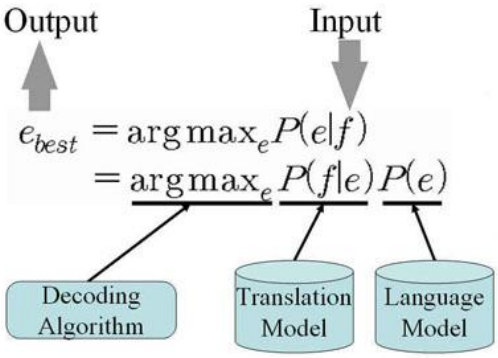

Figure 2.1: The SMT Process (Intelligent Software Lab 2011)

In other words, SMT can be described as an approach whereby translations are achieved by means of statistical models which are derived from analysis of bilingual and monolingual corpora. The quality of the SMT output relies on the quality and size of the aligned bilingual corpora (Forcada 2010, Hearne and Way 2011). The requirement for a large amount of aligned data can be a problem, especially for minority languages and lesser-used language pairs. At the time of writing, the most commonly used corpus in MT research is the Europarl Corpus which was extracted from the proceedings of the European Parliament in 11 European languages: Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish. However, there are instances where corpora can be created artificially, for example, for the Spanish-Chinese language pair as described in Banchs and Li (2008).

Hassan *et al*. (2006) identify phrase-based statistical machine translation (henceforth PBSMT) as the leading research interest within SMT. The introduction of PBSMT (Koehn *et al*. 2003, Och and Ney 2004) marked a considerable advance with regard to translation quality, and many researchers (e.g. Blunsom *et al*. 2008) ascribe the rising popularity of SMT to the advent of phrase-based and syntax-based approaches (see below). The phrase-based approach treats 'phrases' as the basic units of translation, where a phrase can be comprised of any sequence of words that appear in succession but are "not necessarily linguistically meaningful" (Chan *et al*. 2007, p. 33). During the translation process a phrase table is populated with potential translations of a list of phrases derived from the source text; this phrase table is then arranged in a specific order depending on the statistical data gathered from the corpora (Chen *et al*. 2008).

No approach to MT is without its shortcomings, however. SMT, for example, has problems with word reordering (Och *et al*. 1999), and has difficulty translating from morphologically poor languages, e.g. English, to richer ones, e.g. Arabic (Koehn 2005, Avramidis and Koehn 2008). On the other hand, research has ensured considerable progress, and word sense disambiguation systems have been shown to improve SMT translations (Chan *et al*. 2007) as has the alignment of word order between languages, which allows for better exploitation of the bilingual data used to train the SMT system (Nießen and Ney 2001).

While SMT systems automatically 'learn' linguistic rules and information from monolingual and bilingual parallel corpora, this information may need to be coded manually in the RBMT process. Typically, as SMT relies wholly on human translations in its corpora, its output can be more fluent or natural given that it uses the building blocks of the original human translation in its new translation, whereas RBMT systems rely on dictionary and rule information and may translate too literally. Similar to RBMT, SMT systems tend to have difficulties with ambiguous and expressive language e.g. idiomatic expressions. Lastly, vocabulary that is not present in the corpora poses a problem for SMT systems and will usually result in the unknown word remaining in the SL, while RBMT systems may have an advantage in understanding how the word should be treated owing to its incorporation of linguistic information.

In light of such problems, MT researchers have looked to hybrid and multi-engine systems (as introduced above). These hybrid or multi-engine MT systems combine traits of systems described above in an attempt to exploit their positive attributes and overcome their shortcomings. Carl and Way (2003) describe hybrid systems as those that focus on "the integration of relatively autonomous subsystems", whereas multi-engine systems run different MT systems "often implementing different MT paradigms" in parallel to accomplish a translation task (ibid., p. xxi-xxii), e.g. Microsoft's hybrid MT system (Aikawa *et al*. 2001), which consists of rule-based, example-based, and statistical components, and the example-based and statistical system described by Groves and Way (2006) who also note that it is becoming "increasingly difficult to distinguish between EBMT and (phrase-based) SMT models of translation" (ibid., p. 189).

Additionally, multi-engine approaches have been used in recent and on-going research. For example, Eisele *et al*. (2008) present two methods of combining RBMT and SMT approaches: in the first case, RBMT engines are used to enrich the lexical resources available to the SMT decoder; in the other case, parts of the SMT structure are used, together with linguistic processing and manual validation, to extend the lexicon of a RBMT engine.

In sum, it has become evident that distinctions between MT paradigms are not as strict as they once were and such a blurring of boundaries has allowed

combinations of approaches that have led to better results than one approach alone could hope to achieve.

Finally, hybridity also extends to the ways in which MT output is post-edited. Simard *et al*. (2007) show how RBMT output can be post-edited automatically with the help of a PBSMT system. Similarly, Terumasa (2007) uses RBMT with SMT for post-editing for the Japanese-English language pair. While post-editing is intended to improve MT output, controlled language is sometimes used to 'improve' MT input, by making source texts easier to translate by machine. Other examples of automatic post-editing of MT output are described by Dugast *et al*. (2007), and Lagarda *et al*. (2009).

2.4.4 Machine Translation and Controlled Language

Controlled language in the context of MT involves imposing some form of restriction (e.g. in the form of writing and grammar rules, and lists of permissible expressions and terms) on the input before it is processed by an MT system. CL is of interest in the area of machine translation, as a restricted input can, in theory, improve the quality of MT output and thus reduce or possibly eliminate the post-editing effort. An example of a CL used to improve machine translatability is provided by Xerox's application of Xerox Multinational Customized English, which was used in conjunction with Systran (Elliston 1979). Other well-known implementations of CL in MT workflows include the use of Caterpillar Technical English in the KANT RBMT systems (Mitamura and Nyberg 1995), and General Motors' use of CL in METAL applications (Means and Godden 1996). More recently, Ford's Standard Language CL has been used in conjunction with MT (Roturier 2006).

Much of the early research into CL and MT was aimed at establishing the relationship between the use of CL input and the quality of MT output: Nyberg and Mitamura (1996), for example, found that the accuracy of MT output they studied depended heavily on the level of control present in the source. Bernth (1999) examined the impact of a set of CL rules on MT output at IBM, and found an improvement in translation quality. Similarly, Bernth and Gdaniec (2001) found translation improvements when a CL was used to rewrite instructions which were then translated into French, German, and Spanish by different MT systems and evaluated by native speakers of the target language.

Later research also focused on the relationship between CL input and post-editing speed and effort. O'Brien (2006, p. 177), for example, found that "controlling the input to MT leads to faster post-editing". Likewise Aikawa *et al*. (2007) concluded that the use of CL improves post-editing productivity as well as MT quality. O'Brien and Roturier (2007, p. 7) went on to investigate which particular CL rules had the greatest impact on post-editing effort and MT comprehensibility, and found violations of rules regarding spelling, use of the semi-colon, use of question marks in the middle of segments, use of double hyphens, and sentence length, to be the most serious.

Gough and Way (2003), building on the work of Schäler *et al*. (2003) and Carl (2003), filter the *output* of their RBMT system so that it follows a set of CL rules. Other examples of such 'controlled generation' can be seen in Bernth (1998) and McCord and Bernth (1998), who describe the identification of unwanted constructions in the parse tree during the translation process so that the text can be reformulated in a more acceptable way. Likewise, Yamada *et al*.'s (2000) transfer-driven MT system used controlled generation to deal with politeness phenomena in translation from English to Japanese.

2.4.5 Machine Translation Evaluation

Traditionally, the evaluation of MT output has been carried out by human evaluators with linguistic competence who have been trained to some extent to measure concepts such as fluency, accuracy, and overall quality. Fluency is understood here as the extent to which the target text 'reads well' in the target language, and accuracy is understood as the extent to which the target reflects the meaning of the source text. Drawbacks to such a method of evaluation are that it can be resource intensive, and may produce different results from one evaluator to another and even from the same evaluator on separate occasions. Along with the growth in the development of MT systems, a need arose to ascertain if changes to a system resulted in quantifiable improvements. While human evaluation would be an ideal method for such an evaluation, it may simply not be possible, especially as systems may be changed many times in short succession. Therefore, a resource-cheap means of evaluation was sought to assist MT developers in the evaluation of their systems. This was the motivation for the development of automatic evaluation metrics or AEMs. The basic premise of AEMs is that the "closer a machine translation is to a professional human translation, the better it is" (Papineni *et al*. 2002, p. 311). To make this comparison, AEMs are given a reference translation created by a human translator, which is typically assumed to be the 'gold standard' or ideal translation.

The most commonly used AEMs are string-based in that they compare strings of the MT output text to those of the reference translation. String-based AEMs include General Text Matcher or GTM (Turian *et al*. 2003) and Bilingual Evaluation Understudy or BLEU (Papineni *et al*. 2002). Such metrics can be useful for charting the development of an MT system in time, however, AEMs are difficult to interpret outside of the MT research community in that it remains unclear if higher scores on an AEM truly equate to a better translation. Nevertheless, AEMs are in widespread use in MT and they provide valuable information, which is often used for the comparison of MT systems (e.g. Callison-Burch *et al*. 2006, Huang and Papineni 2007). Other notable AEMs in the literature are: Meteor (Banerjee and Lavie 2005) and Translation Edit Rate or

TER (Snover *et al*. 2009), both of which allow for the two strings being compared to differ in the use of synonyms without being penalised and both of which allow for multiple reference translations.

The link between AEM results and human evaluations has been the subject of much debate and research (Coughlin 2003). There is evidence to support the belief that AEMs correlate well with human judgment in certain contexts (Kuleska and Shieber 2004) but not others (Och and Ney 2004). BLEU has been shown to have correlate well with human evaluation at the corpus and document level (Specia *et al*. 2010), although its accuracy at sentence level is thought to be questionable (Callison-Burch *et al*. 2006).

Working on similar corpora to the current study, Tatsumi (2009) found GTM to have a stronger correlation than either TER or BLEU with post-editing speed where a higher GTM score was reflected in faster post-editing. Similarly, Sun (2010) also found GTM to have the strongest correlation with post-editing speed. Once again BLEU and TER were used but showed weaker correlations and it was postulated by Sun that GTM scores are best suited to simple sentences rather than more complex or incomplete sentences.

Other studies have found GTM to correlate best with human evaluation involving European languages, e.g. Cahill (2009), Agarwal and Lavie (2008). In the context of this study, GTM, BLEU and TER are used in the evaluation of the MT output and will be described in more detail in the next chapter.

2.4.6 Section Summary

This section focused on the topic of machine translation. It provided a description of how machine translation systems have developed and the current use of rule-based and statistical machine translation systems employed in the current study, and operationalised in the next chapter. This was followed by an exploration of the use of controlled language in conjunction with machine translation systems, and finally, a review of the evaluation of machine translation systems by means of automatic evaluation metrics.

2.5 Eye Tracking

2.5.1 Section Overview

While a increasing number of studies using eye tracking been carried out in translation studies and related areas of translation process studies, and audio-visual translation, much relevant information is available from earlier eye tracking studies conducted in related fields such as cognitive psychology, psycholinguistics, and usability research. In the following, the literature related to translation studies will first be reviewed as it is most relevant to the present research. This will be followed by a review of the most relevant work from a much larger body of literature from the latter domains.

Fundamentally, an eye tracker monitors and records activity/movements of the eyes, and the pupil in particular, by means of video and infrared cameras. The data gathered by the hardware are processed and made available for examination by supporting software. An example of such a setup is the Tobii 1750 (www.tobii.se) and its supporting software Tobii Studio.

To facilitate comprehension of the following paragraphs, a brief explanation of common eye tracking terminology is provided below:

- Area of interest (AOI): an arbitrary area defined by the researcher and usually intended to coincide with a specific visual or textual phenomenon (e.g. the headline of a text) in the material being examined by participants in an eye-tracking study;
- Fixation time/duration/length: the duration of time the eye focuses on an item;
- Fixation count: the number of occasions on which the eye focuses on an item;
- Gaze time/observation length: the duration of time spent gazing within a particular AOI;
- Pupil dilation/pupil size: the size of the pupil in millimetres and its constriction and dilation in response to stimuli (e.g. external stimulus, internal cognitive effort);

- Regression: “any eye movement that begins at the right-most point the reader has fixated and leaves the currently fixated region to the left' (Pickering and Traxler, p. 945);
- Scan path: the way in which the eyes look at items (e.g. a line of text);
- Saccade: a movement from one point of fixation to another; this movement is not fluid and is typically made in a series of short jumps which humans are unaware of.

2.5.2 Eye Tracking in Translation Process Studies

Eye tracking has been adopted as a research method in translation process studies (e.g. O'Brien 2006, Jakobsen and Jensen 2008, Pavlovic and Jensen 2009, Jensen *et al*. 2009), and has been used as a supplement to keystroke logging (e.g. Dragsted and Hansen 2008, Sharmin *et al*. 2008), functional magnetic resonance imaging (fMRI), e.g. Chang 2009, and in the evaluation of machine translated content (e.g. Caffrey 2099, Flanagan 2009).

O'Brien (2006) tests eye tracking as a methodology in an investigation of translators' interactions with Translation Memory (TM) tools. The subjects comprised four professional translators who were familiar with both the TM software and the text domain. The subjects were required to translate a text from English into their native language (French for two, German for two) using the TM while being unobtrusively monitored by the eye-tracking equipment. The subjects were monitored by the eye tracker as they translated to the various segments containing the different types of matches provided by the TM, i.e. No Match, Fuzzy Match, MT Match and Exact Match. O'Brien uses a reading task as a baseline measure for later comparison of data, and retrospective interviews were also used to provide additional data for the study. Processing speed, measured here as the number of source-text words processed per second, was used to show that Exact Matches were processed faster than other matches, that Fuzzy Matches and MT Matches were processed slower but at a similar speed to one another, and that lastly No Matches required more time to process.

In addition, O'Brien (ibid.) uses eye-tracking technology to measure subjects' pupil dilation, which she states 'can be used as a measure of cognitive effort' (ibid., p. 191). Percentage change in pupil dilation (see Iqbal *et al*. 2005) is adapted for use in this case and the assumption is made that a higher percentage change (in pupil dilation) indicates greater cognitive effort. For the most part, this measurement yields results consistent with O'Brien's findings on processing speed.

Lastly, O'Brien uses the gaze replay function of the eye-tracking software to examine Fuzzy Matches in more detail and shows how this function, especially when used in conjunction with retrospective protocols, can provide useful

qualitative information regarding the experiment design and for later use in data interpretation.

Chang (2009) adopted eye tracking in his study of the effort involved in translating from a native language to a second language and vice-versa. The main experiment in the paper involved sixteen translation students with Mandarin Chinese as their native language and English as their second. Cognitive load was estimated by eye-tracking indicators, including pupil dilation and fixations. Results show that when participants translated from Chinese into English, pupil size was significantly larger than when they translated from English into Chinese, suggesting that translation into a foreign language requires more cognitive effort. This was supported by the higher number of fixations on screen in the Chinese to English translations. Additionally, fixation time was higher for the target text than the source in both directions, and the frequency of fixation within the area of interest in the TT is higher than for the ST.

Jakobsen and Jensen (2008) examined the differences between reading and translation tasks with a group of twelve participants, of whom six were professional translations and six were student translators. The keystroke logging software Translog was used to display the text and record keyboard activity. Using a Tobii 1750 eye tracker, the model used in the current study, they investigated the differences in processing ST and TT in terms of fixation count and duration, gaze time, and attentional shifts between ST and TT windows. They found that all measures were greater, but not statistically so, in the TT window representing greater cognitive effort on that side of the translation process (see below). Lastly, frequent attentional shifts between ST and TT windows were also evident.

Sharmin *et al*. (2008) studied eighteen student translators in their translation of three texts from English (their foreign language – L2) into Finnish (their native language – L1). Like Jakobsen and Jensen (2008), they found fixation duration to be significantly greater for TT processing.

Pavlovic and Jensen (2009) examined translation directionality with a group of sixteen translators, eight professionals and eight students, who translated a single text from Danish (L1) into English (L2) and vice versa. They found that, once again, TT processing is more cognitively demanding than ST

processing, as evidenced by significantly higher gaze time, fixation duration, and pupil dilation when translating into L1. However, when the researchers attempted to differentiate between translation into the native language and into the foreign language, only pupil dilation was significantly higher when participants translating into the L2.

Hvelplund (2011) used eye tracking to investigate the allocation of cognitive resources in the translation process with 24 participants, twelve professionals and twelve students. The study found that TT processing resulted in longer gaze time and supported the hypothesis of parallel processing in translation (see below) whereby translators engaged in simultaneous processing of source and target text. It was also found that professional translators focused more attention on TT reformulation, while students put more effort into ST comprehension. Pupil dilation was significantly larger during the TT reformulation stage than during the ST comprehension stage across both groups, while text complexity (as measured by Flesch and LIX readability indices, word frequency, and non-literal expressions) did not result in differences in pupil size. Time pressure was also shown to affect pupil size and Hvelplund (2011, p. 237) concludes that "this extra workload on working memory is reflected in larger pupils".

Finally, in the area of audio-visual translation, especially for the reading of subtitles, eye tracking has been used as a method of empirical investigation (e.g. d'Ydewalle *et al.* 1991, d'Ydewalle and de Bruycker, 2007, Koolstra *et al.* 1999). Of greater similarity to the current study Caffrey (2009) explored viewers reading of cultural notes attached to subtitles, and the differences between one and two-line subtitles and the viewer's eye behaviour. He also found that fixation length was significantly correlated with cognitive effort across these areas, which adds support the use of eye tracking as a method in tasks involving more than one language.

The above studies are particularly useful in that they identify eye tracking metrics that can be particularly associated with cognitive effort in translation scenarios. They are thus of interest to the current investigation of readability and comprehensibility of MT output. However, the studies reviewed above also problematize some aspects of eye-tracking methods and the issues raised are

worth attending to here. Firstly, a number of studies have identified what O'Brien (2006) refers to as an "acclimatisation" effect in observations of pupil dilation. This happens when "pupil dilation starts at a value that is higher than the average, but adjusts itself quickly to average levels" (ibid., p. 192); this phenomenon has also been observed by Hyönä *et al*. (1995).

Ericsson (2000, p. 248) provides an explanation for an acclimatisation effect in measures of gaze time: "readers can be interrupted by some attention-demanding, yet irrelevant task, and then resume (their previous task) with minor difficulties"; initial gaze time is therefore higher as long-term memory is being accessed. Callicott *et al*. (1999) state that while certain parts of the brain can become more engaged as cognitive effort and demands increase, others disengage once a fixed threshold is reached. Such a threshold varies across individuals and with factors such as expertise, mental state, motivation, and task duration.

Secondly, while some studies show strong correlations between pupil dilation and cognitive load (Hess and Polt 1964, Beatty 1982, Hyönä *et al*. 1995), Iqbal *et al*. (2005) argue that any measure of pupil size has to include a validated task model before any conclusive findings can be made. They (ibid.) find that cognitive load as measured via pupil size decreases more at task boundaries higher up in the model's hierarchy and less so lower down. From the point of view of the current study, however, no such valid task model, or indeed hypothesised model exists for the (human) evaluation of MT output.

Thirdly, other research suggests that pupil responses are not a reliable measure for any one particular stimulus and they should be supplemented with additional measures to ensure valid and reliable results (Beatty and Lucerno-Wagoner 2000). These are issues that are revisited in the next chapter on methodology.

2.5.3 Other Studies

The eye-tracking research reviewed in this section focuses on more generic aspects of cognition, language processing and interface usability. Early eye-tracking research into text complexity found that texts deemed to be conceptually more difficult resulted in an increase in fixation duration, a decrease in saccade length, and an increase in the number of regressions observed (Jacobson and Dodwell 1979, Rayner and Pollatsek 1989), although definitions of such levels of difficulty are not provided. Roughly 10-15% of all fixations are accounted for by regressions, and evidence suggests that many of these are due to comprehension difficulties (Just and Carpenter 1980, Frazier and Rayner 1982, Hyönä 1995).

In research related to eye-tracking studies of translation, Hyönä *et al*. (1995) examined 'pupillary response as an independent on-line measure of cognitive load' (ibid., p. 598) in language processing tasks in interpreting. The first experiment in this paper concerns an examination of pupil size over nine subjects carrying out simultaneous interpreting (English to Finnish, where all subjects were native speakers of Finnish, fluent in English and familiar with simultaneous interpreting) and other less demanding language tasks, in this case listening to and repeating back or 'shadowing' a message. The average pupil size during these tasks was measured and the results indicated that pupil dilation increased with task complexity. This experiment left the authors with the concerns that: a carry-over effect of task difficulty may have occurred, the results could reflect an increase in the level of general arousal, and output requirements have an influence on the results. A second experiment was devised to address these issues and was performed using single words instead of passages of text. The experiment was modified to include eighteen subjects and a listening task, where subjects were required to say "yes" aloud after they had recognised the word; in shadowing they were to repeat the word aloud; and the interpreting task was replaced with "a lexical translation task in which subjects were to give a meaning equivalent for the word in the output language" (ibid., p. 604). The findings of this second experiment mirror the results of the first and show that words which were chosen for their increased difficulty caused higher levels of

pupil dilation across all observed tasks, as did repeating back words in a non-native language. It was found that "after a response had been given to a target word, the pupil constricted back to the baseline level" (ibid., p. 610). This is consistent with Beatty's (1982, p. 288) claim that "the effects of emotional arousal are generally longer lasting than the brief phasic responses evoked by cognitive activity", the latter being apparent in the above study. In addition, it was noted that an increase in pupil dilation was still observed as a carry-over effect. Overall, the findings show how pupil dilation can be employed as an indicator of cognitive load during language processing.

Hyönä and others have also carried out other research in this area using eye tracking to examine lateralised word recognition (Hyönä and Koivisto 2005), and text comprehension (Kaakinen and Hyönä 2005), and to investigate differences in reading styles among adult readers who upon questioning were largely aware of their reading style; the findings contribute to the authors' understanding of fast and slow linear readers (Hyönä and Nurminen 2006).

Andersson *et al*. (2006) combine eye tracking with keystroke logging to study writing processes in subjects of different ages and writing skill. Their work serves as an example of how eye tracking can be used to supplement already existing methodologies to provide more reliable empirical results. In this case, some of the limitations of keystroke logging are overcome by the use of eye tracking in that the subjects' eye movements can be monitored even during periods of keyboard inactivity, which in turn can be due to cognitive processing. The authors use the ScriptLog keystroke logging program and a head-mounted eye-tracking system, and create a fixed experiment area to ensure comparability and consistency in the data and to provide the eye tracker with fixed planes to monitor, i.e. the keyboard, the monitor, and the source text displayed on a separate sheet of paper. In their explanation of the methodology, the authors warn that "an eye tracker covers approximately 60 degrees of visual angle for a writer who does not turn his/her head. If the writer looks further away, data will be partially lost" (ibid., p. 3). The authors give an example of the usefulness of the combination of video recording, eye tracking, and keystroke logging from their experiment and highlight the need for further study in the area.

Bartels and Marshall (2006) used eye tracking as a means of understanding their subjects' behaviour in experiments designed to test the accuracy of several human performance models on a single complex task. According to the authors, "cognitive models can be used to identify a particular visual pattern as evidence of a specific cognitive strategy" (ibid., p. 142), and these models can be validated or modified by eye-movement data. The experiment focused on in this paper saw fourteen carefully selected participants complete an air traffic control task designed to elicit eye-movement and cognitive-process data. The eye-tracking setup consisted of small video cameras mounted on a lightweight headband and the task screen was separated into seventeen regions to cover action buttons and message windows etc. Particular attention was paid to the total viewing time spent in each of these regions and the number of transitions between them. Participants completed a follow-up questionnaire and retrospective interview. Results showed differences between the two types of display conditions tested in the experiment, i.e. colour and text displays, and subsequently showed different strategies being used to complete the task. The analysis of the data gathered by the eye tracker allowed the authors to develop an explanation of the strategies adopted in the different scenarios and serves as an example of the usefulness of eye tracking in human cognition studies.

Nakayama *et al.* (2002) examined 'oculo-motor indices' which denote measurements of pupil size, blink rate, and eye movement. In their experiment, pupil size and blink rate increased with task complexity. Conversely, the occurrence of saccades and their length decreased as task complexity. The oculo-motor indices were found to respond to task difficulty and the authors deduced that pupil dilation and eye-movement frequency are indicators of cognitive workload.

Hess and Polt (1964) carried out an experiment where subjects solved simple mathematical problems while the authors used eye tracking techniques to record and examine exact changes in subjects' pupil size. Five subjects participated in this experiment. The experimenters began by asking each subject to focus on the number 5 prior to the mathematical tasks and a gap of 30 seconds was allowed between each task. This allowed the participants to adjust to their

new surroundings and possibly avoid the acclimatisation effect referred to in section 2.5.3 above. The research found that "there is a tendency for later stimuli in a series to get a slightly smaller response than earlier responses in a series if the stimuli are equal in value" (ibid., p. 1191). The authors adopt a method of increasing task difficulty in their experiment to avoid this effect. Additionally, it was found that subjects' pupils showed a gradual increase in size and reached their recorded maximum immediately before their responses, after which the pupil size reverted back to the measured control size. The authors came to the conclusion that there is a direct link between cognitive activities and pupil responses.

Holsanova *et al*. (2006) use eye tracking in a five-subject experiment to test general assumptions regarding reading of newspapers. They examine so-called 'entry points' and 'reading paths', which show respectively where readers first look and their gaze paths. The results help to establish different types of readers based on their reading patterns: editorial readers, overview readers, and focused readers.

Other studies have examined eye tracking measures in tasks such as bilingual reading where, for example, Altarriba *et al*. (1996) found that the increased cognitive demands of reading bilingually resulted in shorter fixations, longer saccades, and fewer regressions. A great volume of eye-tracking research has been carried out in usability studies, where the topics of cognitive effort, memory, and ease of use of content typically arise. Such usability studies represent investigations of fundamental aspects of cognition that are central to the current study.

Goldberg *et al*. (2002) adopt eye tracking in their research into the usability of a prototype web portal application in an effort to uncover more about how users perform certain tasks with the software and so influence and improve future web design. The authors classify eye-tracking analyses into two types: "top-down" analyses, which start with a cognitive or goal-driven model and use the derived data to prove or disprove aspects of the model, and "bottom-up" analyses, which "attempt to develop behavioural inferences, starting from model-free eye-tracking derived data" (ibid., p. 51). In their experiment seven participants were asked to perform six specific tasks while eye-tracking

equipment monitored their behaviour. The findings show that what is referred to here as 'localized learning' occurs; this is said to be a result of mental representations of the location of interface features and results in shorter scan paths with fewer fixations and saccades. Localized learning is also evident in the decrease in fixation durations observed as the tasks progressed in a fixed order.

In a study of problem solving, Rudmann *et al*. (2003) test sequential cognitive models by asking participants to follow animated gear movements onscreen. They test the hypothesis that cognitive activity is related to the object of fixation by interrupting subjects' problem solving mid-task and asking them to state what they were thinking about. The researchers found that in most cases the hypothesis is confirmed.

In Eger *et al*. (2007) 24 subjects completed a search task and data were elicited using think-aloud protocols (see section 2.6.5), eye tracking, and 'retro eye cue' (retrospective protocols cued by eye-tracking data). It was found that with the latter method participants identified the greatest number of usability problems, thus demonstrating the value of eye-tracking data in usability research.

As this brief overview has demonstrated, eye-tracking studies have yielded useful results in a variety of areas, but other reviewers such as Alves *et al*. (2009) have raised questions about the reliability of certain measures e.g. fixation count and duration. Such questions are addressed in the chapters five and six in relation to the current study.

2.5.4 Section Summary

This section focused on the topic of eye tracking in several research domains. It first described eye-tracking studies conducted in translation process research and highlighted several inconsistencies of findings and the importance of strict experimental controls such as environmental aspects and consistency of eye-tracking measures.

The review was then expanded to include other in the areas of cognition, psycholinguistics, and usability, which demonstrate that such problems are not unique to the use of eye tracking in translation process research, and therefore, require careful consideration for use in research projects such as the current study. Further descriptions of the eye-tracking method used in the current study are provided in the next chapter.

2.6 Cognitive Aspects

2.6.1 Section Overview

This final section concerns the cognitive aspects relevant to the study. Firstly, in sub-section 2.6.2, human memory systems will be described and the topics of memory decay (2.6.3), recall (2.6.4), and automated processing (2.6.5) will be addressed. This is followed by a discussion of Think-Aloud Protocols (2.6.6), a methodological tool used to elicit data from participants' verbalisation of cognitive activates during experiments. The section concludes with a presentation of the cognitive framework relevant to the current study (2.6.7).

2.6.2 Human Memory

Since James' (1890) coining of the terms 'primary' and 'secondary' memory stores, there has been an ongoing debate on the topic of memory. The terms 'short-term memory' (STM) and 'long-term memory' (LTM) are, however, commonly used. STM refer to memories that are in conscious awareness and currently being attended to. LTM refers to memories that are not in conscious awareness but are stored awaiting recall when required. The distinction between STM and LTM is referred to as a dual-store theory, of which others exist (see Eysenck and Keane 2010). Evidence has led to the concept of dual-store being largely accepted and various hypothetical models of memory structure have since been proposed. (e.g. Scoville and Milner 1957, Baddeley and Warrington 1970)

One such model was Atkinson and Shiffrin's (1968), which includes a preliminary stage where sensory information is held in the sensory store (see below in this section) before being processed. Multi-store models were popular for many years (Healy and McNamara 1996) but in recent times, the model of working memory has dominated.

Earlier models regarded STM and LTM as two separate memory stores with difference in capacity and duration. Baddeley and Hitch's (1974) model of working memory (henceforth WM) postulated the STM as an active workspace for a variety of processing tasks including the processing of new or old memories. In this model the LTM is still seen as a passive and larger memory store, which consists of declarative memory and procedural memory. Declarative memory consists of semantic memory, which stores factual information, and episodic memory, which stores autobiographical memories. Procedural memory holds information about series of actions frequently preformed (e.g. with motor skills) and, once stored sufficiently, is activated under the level of conscious awareness.

Baddeley and Hitch (ibid.) subdivided WM into the central executive (a modality-free administrative component similar in function to attention) and its slave systems of the phonological loop (which holds information in phonological form) and visuospatial sketchpad (for spatial and visual coding), and later

Baddeley (2000) added the episodic buffer (which links information to form units e.g. visual or verbal sequences). Figure 2.2 illustrates each of these components.

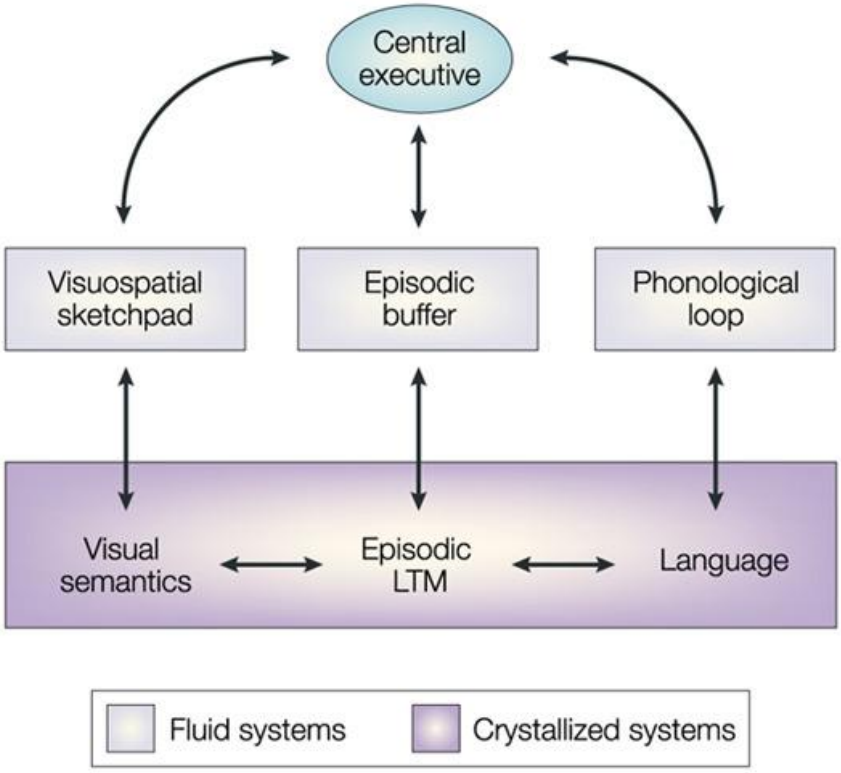

Figure 2.2: Baddeley's (2003) Revised Multi-Component Working Memory System

Sensory stores hold information for a very brief time and are specific to each main sense: iconic (visual), echoic (auditory) and haptic (touch). As the current study is concerned with visual stimuli, there is a brief description of iconic memory which was presented by Sperling (1960), who claimed that information is stored for up to 250ms after the offset of the stimulus. The role of these sensory stores is to provide a buffer for information to be processed further.

2.6.3 Memory Decay

Early work on memory decay was pioneered by Ebbinghaus (1913) who postulated two main theories to explain why memory decay occurs: spontaneous decay (memories deteriorate with time, regardless of other input), and interference (memories are disrupted by the influence of other input, usually more recent).

A revised version of the decay theory was proposed by Thorndyke (1921), who suggested that memories decay if they remain unretrieved. Bjork and Bjork (1992) call this the decay with disuse theory and suggest that it is not the memory itself that deteriorates but access to it. Frequent retrieval is therefore necessary to ensure the retrieval routes are kept. Ebbinghaus (ibid.) showed the 'forgetting curve', a rate of forgetting which was extremely rapid at first but over time became more gradual; this has since been supported by more recent research (e.g. Slamecka and McElree 1983). It is now accepted that newly learned material is forgotten at a rapid pace initially, but then this rate becomes more gradual over longer periods of retention and recall.

Adding meaning to an item makes it easier to remember as is evident from the use of mnemonics. Such techniques use items already in a person's LTM store and associate the new items with this already stored knowledge (Bransford and Johnson 1972). This is added to by Craik and Lockhart (1972) who propose the 'levels of processing' theory, which suggests that the processing of new input involves the extraction of information at various levels of increasing depth.

Craik and Lockhart (1972) propose three such levels of processing: structural, acoustic, and semantic. They show that participants who were forced to employ a deeper level of processing had better retrieval than participants in tasks where only shallower processing was required and these findings have been supported in other work (Hyde and Jenkins 1973, Parkin 1983). The theory of different levels of processing was added to by Craik and Tulving (1975) who replaced the sequential model with a parallel model known as the elaborative encoding theory and which assumes all new input is processed in the three different ways simultaneously. Despite initial criticism, this theory is still accepted in contemporary research.

Repetition of an item alone is not sufficient to retain it properly, as demonstrated by Craik and Lockhart (1972), who distinguish between maintenance rehearsal such as repeating a word aloud, and elaborative rehearsal, whereby the item is processed more deeply. They argue that the latter would lead to storage in the LTM, whereas the former would only hold the item in the conscious awareness of the STM where it would be forgotten after a short period of time unless it was processed further. Evidence such as that in Craik and Watkins (1973) shows that repetition of an item with no deeper processing does not lead to storage in the LTM. Another example is given by Nickerson and Adams (1979) who demonstrate how participants are unable to remember details of coins they have used on a daily basis for a long period of time. In addition to this, studies have shown that participants tend to have higher recall scores when they have used reference to themselves, i.e. the self-reference effect (Klein *et al*. 1989, Kahan and Johnson 1992). It has also been found that participants have increased recall of items when they themselves have generated the items in a word-association text (Smith and Healy 1998)

Tulving (1972) proposes that item retrieval depends on the availability of retrieval cues that match aspects of the stored memory. This encoding specificity principle proposes that retrieval cues will only be successful if they contain some of the same specific items of information that were encoded with the original input. It has also been proposed that the probability of retrieval depends on the amount of feature overlap between the two items. This builds upon the aforementioned concept of deep processing, i.e. the deeper the processing of an item, the more associations are made and so the greater the likelihood of later retrieval via a matching cue.

Greenspoon and Ranyard (1957) propose that the context of retrieval is of importance. Their participants were tested for recall in two groups, one in the room where the material was learned, and the other in a different room. They found that former group had better recall, a finding supported by other studies (Godden and Baddeley 1975, Jerabek and Stading 1992).

It is commonly found that participants recognise far more items than they can recall, while cued recall tends to yield results somewhere between free recall and recognition (Tulving 1972). Most recall tests available are predominately

used to test LTM in the form of recall of wordlists or stories. Tests of STM can be more difficult to devise, but one of the more popular is the testing of immediate memory span. There is, however, evidence that LTM makes a contribution to performance in this test (Hulme *et al*. 1991).

A final point to note here relates to what is known as the recency effect, which has shown that participants are more likely to remember items from the end of the list than from the middle (Eysenck and Keane 2008). This is due to the later items still being held in STM, whereas earlier items will have been lost unless they have been stored in LTM, which would have not been possible as the participants are still processing new information. It has been found that by introducing a delay between learning, in this instance, a wordlist, and recalling it, the recency effect is neutralised (Glanzer and Cunitz 1966).

2.6.4 Memory and Recall

Scholars of memory distinguish between encoding (input), storage, and retrieval (output). All three stages need to be successfully completed before a memory can be properly retrieved. An inability to recall an item could be due to a failure at any point in the process, for example, lack of attention at the input stage.

There are three main methods of testing memory performance:

- Free recall (or spontaneous recall): participants are asked to generate test items from their own memory without outside interference at that given time;
- Cued recall: participants generate items as above but with the assistance of cues which help to remind them of an item or in other words jog their memory;
- Recognition: participants are presented with test items and asked if they recognise them.

Working memory span tasks, like the model of working memory itself, are widely used in psychology and related disciplines. Conway *et al.* (2005) believe that WM span tasks predict cognitive behaviour across domains, including reading comprehension and problem solving. They stress the functional importance of the WM and highlight the need for WM span tasks to measure storage and rehearsal but also simultaneous processing of other external task-specific information such as word or sentence comprehension.

The reading span task (Daneman and Carpenter 1980) consists of a word span task with the additional task of sentence comprehension. Participants read sentences of two to six words once for word retention. Word recall is promoted on completion of each grouping, although there are several variations of this task (for a full review see Conway *et al*. 2005). Conway *et al*. (2005) come to the conclusion that the best measurement includes scoring procedures that "exhaust the information collected with a task" (ibid., p. 776) and partial-credit scoring. In other words, correct responses to items within a group are assigned a number

with partial recall being rewarded on a proportional basis. Additionally, Moravcsik and Kintsch (1993) showed how more skilled readers consistently perform better in recall than less skilled readers. Schneider *et al*. (1990) show that general intelligence is unrelated to text recall.

WM is considered to be of limited capacity and early work (Miller 1956) suggests a span of seven elements or 'chunks' regardless of whether these chunks are letters, words or numbers etc. Pascual-Leone (1970) expanded upon this and showed that the span for digits is larger (seven), whereas for letters and words it is smaller (six and five respectively). Letters and words are also dependent on both the time required for participants to speak their contents aloud to themselves, and on the lexical complexity and familiarity of the items.

Finally, some studies have shown links between eye tracking metrics and recall. It has been found, for example, that fixation duration time is positively correlated with recall (e.g. Hollingworth and Henderson 2002). It is also commonly assumed that fixations are required for the extraction and transfer of visual information into the various memory stores and for later recall (e.g. Luck and Vogel 1997).

2.6.5 Automated Processing

Automated processes are generally considered to be those operations which are maintained without conscious control and which require no or few processing resources (Anderson 2000, Eysenck and Keane 2008). Processes are often automated as a result of task repetition and are drawn from procedural LTM. These tasks require intentional initiation in some form, but their continuity is supported by subconscious processing since a translator, for example, does not have to consciously allocate processing resources to maintain the activity. In translation, it can be expected that automatic processing takes place. Processes which involve orthographic analysis during ST reading and the mechanical operation of TT typing do not require conscious processing. Although the translator intentionally initiates the reading process by moving their eyes to the location of the word, visual exposure to the letters activates an orthographic processing stream and an automatic bottom-up processing stream which cannot be interrupted (Valdes *et al* 2005, p. 279). Jääskeläinen and Tirkkonen-Condit (1991) have found evidence that professional translators engage in more automatic processing than student translators (they used speed as a measure of automaticity). However, Dragsted (2004, p. 47) argues that the translation process is an inherently non-automated processed in that it always involves activation of WM, and that the translator constantly has to construe the meaning of the ST or reformulate it in the TL.

A growing pool of evidence exists to support the proposition that reading is not a wholly automatic process (see Rayner *et al*. 2001 for an extensive review). Valdes *et al*. (2005) report that the processing of words is assumed to be automatic once the reader decides to read. They conclude that word processing occurs at several levels and in a bottom-up manner that has been automated (Neely 1991). However, other studies in semantic priming find that the priming does not to be conscious to the reader for it to be effective (see Valdes *et al.* 2005 for a review). Such findings present interesting points for the current study, especially in terms of information available to participants in terms of verbalised thoughts during the Think-Aloud Protocol. They (ibid., p. 293) conclude that "semantic properties of a word can be processed when words

are presented under an objective threshold of awareness, and also when attention is not allocated to semantic but to low-level features of the word" therefore, processing words while reading can be said to be an automatic process but this "should not imply that no control mechanism can operate".

2.6.6 Think-Aloud Protocols

Think-Aloud is a data elicitation technique in which participants are asked to verbalise their thoughts as or after they perform some task. These verbalisations are recorded in a Think-Aloud Protocol (TAP). TAPs were adopted from psychology where they were developed primarily by Ericsson and Simon (1980, 1987, 1993). In early translation process studies, TAPs were the most common type of data elicitation technique but they have since given way to technologically more advanced methods such as eye-tracking, brain imaging, and keystroke logging.

Concurrent TAPs (e.g. Jääskeläinen and Tirkkonen-Condit 1991, Jääskeläinen 1999) are protocols where participants verbalise thoughts *during* the process under investigation. However, this method has been found to have an impact on the process being investigated. Jakobsen (2003, p. 78-79), for example, found that the use of concurrent TAPs significantly increased the time it takes to produce a translation. He also observed differences between professional and student translators, whereby the former made far fewer verbalisations, which suggests that the process of translation, or at least sub-processes thereof, has become automated in professionals and are therefore not available to introspection and verbalisation among this group (Jakobsen 2003). Likewise, Broadbent *et al*. (1986) state that implicit knowledge is often non-verbal and thus difficult to articulate, leading to greater distortion in TAPs than is the case with explicit knowledge, which is encoded, for example, in verbal rules used to solve problems in a given task.

Gile (1998) has furthermore commented on the difficulty student translators have in producing a translation and simultaneously verbalising their thoughts. He cautions that the reliability of TAP data is negatively affected by the allocation of cognitive resources during such simultaneous exercises, as the translation task is already cognitively demanding for participants, and the TAP method puts an even greater burden on them, changing the very process it is supposed to elucidate.

One alternative to concurrent verbalisation or thinking-aloud is to report on thoughts retrospectively. Hannu and Pallab (2000) compare concurrent and

retrospective TAPs and find that concurrent verbalisation provides more insight into the steps leading to a decision while a retrospective approach provides more detail on the decision.

Some researchers have thus sought to overcome the shortcomings of (retrospective) TAPs by triangulating data elicited in this way with data elicited in other ways, and in particular using eye trackers (see Jakobsen 2003, Alves *et al*. 2009). Indeed, Alves *et al*. assume that "retrospection, carried out as free recall and subsequently as guided recall supported by eye-tracking recordings, offers a promising avenue to tap into translators' meta-cognitive activity" (ibid., p. 270). This is mirrored in psychology where introspection, from which TAPs were derived, has experienced a similar journey. Rosenthal (2000) concludes that introspection does not accurately represent mental states nor does it provide insight into concurrent states in any variety of situation, it therefore requires additional supplementary methods to be useful. This is echoed by Lashley (1958) who states that introspection only makes the results of mental processes accessible, not the processing itself.

In light of such misgivings about TAPs as a data elicitation technique, the current study also relies on triangulation of data gleaned from retrospective TAPs with data elicited from other sources, in this case eye tracking. The methodology adopted will be described in more detail in the next chapter.

2.6.7 Cognitive Framework

Building upon the previous sections on aspects of memory and recall, this section establishes a broader description of the cognitive framework vis-à-vis the translation process, and the related activities of comprehension and production, all of which are integral parts of the current study.

Translation and its evaluation, and indeed, the reading process in general, is essentially an information processing task (e.g. Newell and Simon 1972), and a study of such processing should therefore drawn upon established models and research in the relevant areas of cognitive psychology and psycholinguistics.

Cognitive processing has been described under three main paradigms (see Eysenck and Keane 2008 for a review): symbolism (use of arbitrary symbols as representations); functionalism (descriptions of cognition based on the functionality of the components in question); connectionism/parallel processing (distributed networks of neural connections that form patterns of activation sequences as output given a certain input in a given task); and situational and embodied (focus on the effects of the environment and physical attributes of the body and the interaction between the two). These paradigms have shaped the theories and models described below where there are evident developments in terms of the sophistication and empirical support.

2.6.6.1 The Translation Process

Great similarities between the above processes of comprehension and production and the proposed models of the translation process, especially as the latter have been developed from the former. As described by Toury (1985, p. 18), translation processes "are only indirectly available for study, as they are a kind of 'black box' whose internal structure can only be guessed, or tentatively reconstructed. Insight into this black box has stemmed from empirical research into the cognitive aspects of translation, which dates back to the 1980s and has relied heavily on works from cognitive sciences and psychology (Shreve and Koby 1997, p. xii).

Research on language comprehension and production has also developed into central parts of contemporary models of translation (Kintsch 1988, Padilla *et al*. 1999, Anderson 2000). For example, Kintsch's (1988) framework for language comprehension, the construction-integration model, has been applied to the process of comprehension in translation (Padilla *et al*. 1999).

Many theories exist as to what occurs during the translation process. Some definitions of the translation process are more concise than others, e.g. Hansen (2003, p. 26) describes the translation process as "everything a translator must do to transform the source text to the target text", while others are more detailed and focused on cognitive aspects. A commonality to these proposed models is that they have a series of stages.

Shreve and Koby (1997, p. xi) describe the translation process as separate sub-processes: (1) comprehension and interpretation of the SL, (2) transposition of the SL into the TL, (3) and expression in the TL. They highlight the linguistic and sociocultural knowledge that is drawn from the LTM into the STM while the text is being processed at each stage.

Gile (1995) proposes a sequential model of translation, which consists of two phases: comprehension and reformulation. Both phases rely on linguistic and world knowledge. During the process, the translator constructs tentative hypotheses for the meaning of each ST unit (words, phrases, sentences depending on ability etc.). Each of the hypotheses is tested and modified or rejected until the most plausible hypothesis is accepted. This process is repeated for the creation of the TT. Gile (ibid.) states that this model assumes that translation is equivalent to a combination monolingual comprehension and production.

Additionally, the model proposes a serial approach where ST comprehension occurs first, then the production of the TT unit, and repeated over and over until completion. There is evidence to support that translation is not a serial process (e.g. Ruiz *et al*. 2008), however, Gile's model provides an interesting parallel with the development of other models of cognition where earlier models saw cognitive processes as being sequential and later developments moved to parallel processing, and later still embodied approaches e.g. artificial intelligence.

Danks and Griffin (1997, p. 166) propose that comprehension in translation differs from monolingual comprehension in that translation involves more than finding a way of converting ST units into the TT and also introduces factors such as the original individual's intent, the translator's intent, the target audience of the translation and the 'situation model'. The overall model encompasses both top-down and bottom-up processing, and represents simultaneous parallel processing in that the translator can move between ST comprehension and TL production and mix both aspects of the tasks in the overall process of translation. This model has been supported by evidence from reading experiments by Jakobsen and Jensen (2008: 109-111) and shows that reading for translation is more effortful than reading for normal comprehension.

Mossop (2003) proposes three stages of translation: (1) pre-drafting (which takes places before sentence-by-sentence drafting begins); (2) drafting (which involves composition of the translation); and (3) post-drafting (i.e. evaluation). Similarly, Jakobsen (2002) describes the translation process in four stages: orientation (initial comprehension); drafting stage (creation of TT text); end revision (revising TT text); monitoring (evaluation).

Support can be found for a parallel approach to the translation process whereby comprehension of the ST and reformulation of the TT can occur in parallel, i.e. linguistic and world-knowledge of both languages can be accessed simultaneously (Gerver 1976, Ruiz *et al*. 2008). Such propositions contrast sequential views of translation such as Gile's model (e.g. Gile 1995, Seleskovitch 1976). More recent approaches adopt a hybrid view incorporating both serial and parallel models depending upon contexts such as the translator's language competence as a bilingual, experience and domain knowledge (Paradis *et al.*, 1982, Paradis 1994). Several studies have demonstrated empirical support for parallel processing (Isham 1994, De Bot 2000, Dijkstra *et al.* 2000a, Dijkstra *et al.* 2000b, Van Hell and Dijkstra 2002, Hvelplund 2011). However, as hybrid approaches incorporate parallel processing, whereby less experience translators engage in serial processing, such empirical support does not disprove hybrid models (Ruiz *et al.* 2008).

Overall, while differences between proposed models are evident, clear overlapping and paths of developments can be seen. As described in the previous

section, the evaluation task of the current study involves participants comprehending translated text, producing (in their mind) alternative translations where deems appropriate, and ensuring they are satisfied with their new choice over the original translation or other alternatives they proposed. Such processes are greatly similar to those pertaining to the translation processes described in this section and provide a framework for interpretation of results (see Chapter 6) as well as avenues for future research (see Chapter 7).

2.6.6.2 Comprehension and Production

As participants of the current study are asked to evaluate the MT output, the process of comprehension is a key aspect of the cognitive process involved, and therefore warrants further description here.

Kintsch's model (1988) proposes a model of comprehension that consists of five stages. The first two of which are part of the 'construction' aspect of the model, and the remaining three are part of the 'integration' aspect. It describes comprehension as: (1) construction of relationships between words in the text; (2) relating these links to related links from LTM; (3) the more probable interconnected links are selected; (4) the relevant textual representations are stored in what Kintsch calls 'episodic text memory'; (5) the representations are then stored in the LTM for later use.

The model proposed by Padilla *et al*. (1999, p. 63) has five stages: orthographic or phonological analyses of the sensory input, this level of processing precedes actual comprehension; lexical and semantic analyses are performed, during which a meaning of the word is identified; segmentation of the text or discourse is carried out, in which propositional relationships are formed between the words; a propositional structure of the identified propositions is created which draws on LTM; a higher level representation is constructed which involves the elimination of propositions of lesser importance.

Further to these, Anderson (2000, p. 389) describes a model of three stages: perceptual processes (decoding visual information – reading), parsing (construction of meaning via semantic, syntactic, phrasal etc. analyses), utilisation (acting upon the newly obtained information).

Aspects of parallel processing During ST comprehension the translator engages in lexical analysis in order to identify the meaning of an ST word (Padilla *et al*. 1999, p. 63). This involves the phonological loop of WM and LTM. There is also evidence to suggest that potential TT equivalents of ST words are identified in parallel with this process (Ruiz *et al*. 2008, p. 491), and that syntactic processing of the TT occurs at an early stage in parallel with ST comprehension (Jensen *et al*. 2009, p. 331).

In addition to these, Hvelplund (2011) further distinguishes between ST reading and ST comprehension where ST reading is the perceptual decoding of text involving SM, while ST comprehension involves the extraction and reconstruction of the meaning derived from the ST and draws on WM and LTM. Such a distinction is also made in other works (e.g. Kintsch 1988, Danks and Griffin 1997, Padilla *et al*., 1999, Anderson 2000).

Although participants in the current study are not producing explicit forms of translations, it is assumed that given the task of evaluating the translation output, participants will propose and reject translations, where they deem it necessary, until they are satisfied that they have reached the most appropriate translation. Other possibilities are, of course, that they may accept the output as is, not formulate alternatives, or give up their hypotheses at a certain point. It is therefore of interest to describe TT processing and production.

Kellogg's (1996) model of monolingual text production has three groups which include two additional sub-processes: (1) formulation (planning and translating). During planning the individual will construct a pre-verbal message that corresponds to the idea that is to be communicated, these ideas are retrieved from the LTM (Olive 2004, p. 32). In other words, during the planning the individual plans the goals and ideas lexically and syntactically in the mind. (2) Execution (programming and executing), during execution the individual programs and instructs the motor systems to execute the writing event. (3) Monitoring (comprising reading and editing) during which the individual reads the text and performs edits (Kellogg 1996).

All three processes of Kellogg's model involve the central executive, the phonological loop and the visuospatial sketchpad. During formulation, planning involves the central executive and the visuospatial sketchpad in the creation and

organisation of ideas. Translation (i.e. encoding) relies on the phonological loop as well as on the central executive in translating the ideas semantically and syntactically (Olive 2004, p. 35). Lastly, monitoring also involves the phonological loop and the central executive (ibid., p. 62).

In terms of the evaluation process, two types of TT reading are proposed by Hvelplund (2011): reading of emerging TT output and reading of existing TT output. Both types of reading indicate that the translator is engaging in the verification of the TT output as part of the reformulation process. TT reading, unlike ST reading, is not a precondition for translation; a translator is free to translate without ever glancing at the TT, which in essence makes TT reading a facultative process. Hvelplund (ibid.) also states that ST and TT processing overlap. While the current study does not require evaluators to write or speak any proposed alternative translations, it can be argued that such processing occurs when the evaluation calls for an alternative as deemed necessary by participants. Such alternatives must also be evaluated as they are thought of, or indeed, after they are proposed in the mind of the evaluator (as the study does not require or allow them to be recorded elsewhere, e.g. written or spoken aloud.

2.6.7 Section Summary

This final section dealt with the cognitive aspects relevant to the study. First of all, human memory systems were described to establish a context for the methods described later in the study. This was followed by the topics of memory decay, recall, and automated processing, all of which are of interest to recall testing as employed in the current study. Following this, a discussion of Think-Aloud Protocols was presented, which highlighted the need to use this method in conjunction with other more objective measures to ensure validity of results. The section was concluded with a description of the cognitive frameworks relevant to the current study which contextualised the earlier information presented in this section overall.

2.7 Chapter Summary

This chapter reviewed literature relevant to the pilot and main studies presented in this thesis. It commenced with an exploration of readability and went on to focus on the two indices used in the current study, namely Flesch and LIX. A review of related studies helped to establish the link between readability and comprehension. This was followed by an investigation into controlled languages where the commonalties of several controlled languages were described and related studies reviewed. A brief overview as then given of contemporary machine translation, in which both rule-based and statistical machine translation systems were described. The discussion then moved on to the use of controlled language in conjunction with machine translation and the evaluation of machine translation systems. Following this, the eye tracking literature was reviewed and particular attention was paid to translation process studies. Lastly, cognitive aspects relevant to the study were discussed. This included human memory systems, memory decay, and recall. The use of Think-Aloud Protocols as a data elicitation technique was also discussed, and the case was made for using TAPs in conjunction with other methods.

Part II:

Methodological Considerations

Chapter Three:

Methodological Considerations

3.1 Chapter Overview

This chapter presents the methodology used in the research described in this thesis. As no standard methodology exists for a study of this kind, it was necessary to combine and draw upon several approaches to develop an appropriate means to proceed. Given the exploratory nature of the study, issues of validity and generalisability arose. In order to address these issues and to ascertain the suitability of the proposed methodology, a pilot study was conducted (Doherty and O'Brien 2009, Doherty *et al*. 2010), the results of which are discussed in Chapter Four.

The current chapter presents the methodological considerations shared by both the pilot and the main study. Areas of divergence between the two studies are highlighted in Chapters Four and Five respectively. The chapter is structured as follows: firstly, the underlying philosophical approaches of the study are outlined and supported by a justification (sub-section 3.2). Sub-section 3.3 presents the theoretical framework adopted in this research which includes the research questions and hypotheses, the operationalsation of readability and comprehensibility, methods and sampling used, and a discussion of issues of validity and reliability. Sub-section 3.4 describes the concept of readability as it is operationalised in this study; sub-section 3.5 presents the concept of comprehensibility, again as it is operationalised in this study. Sub-section 3.6 covers additional factors that are essential to the study; namely reader type, motivation, domain knowledge, and time constraints. The corpora used in the study are described in detail in sub-section 3.7, as is the controlled language rule set and its implementation (sub-section 3.8). The MT system is described in sub-section 3.9 along with the automatic evaluation metrics used in the study. Lastly, the use of eye tracking in the current study is clarified in sub-section 3.10 where the hardware and software used are described, and definitions are given for the metrics used in the study.

3.2 Philosophical Stance

3.2.1 Approaches

Several philosophical worldviews or paradigms are relevant to this study, namely post-positivism, constructivism, and pragmatism. Post-positivism is commonly associated with the use of quantitative methods in a top-down approach whereby researchers attempt to validate theories and knowledge by means of determination, reductionism (reducing a concept, theory etc. into smaller, more observable and quantifiable components) and empirical observation. Constructivism is associated with qualitative methods of research and a bottom-up approach that aims to understand phenomena via participants and their subjective experiences and views, with the aim of building theories. Lastly, pragmatism is most appropriately related to mixed-methods research and focuses on the question being asked and the methods of answering this question and thus can be "pluralistic and oriented toward what works in practice" (Creswell and Plano Clark 2007, p. 23).

Crotty (1988) describes the epistemology of pragmatism as one of practicality whereby the researcher uses the most appropriate and effective methods to address the research question. With regard to axiology, multiple stances can be held in terms of the biased and unbiased perspectives adopted and tested as well as approaches from other worldview paradigms. Likewise, Tashakkori and Teddlie (2003) argue that both quantitative and qualitative methods may be used in a single study, and that the research question is more important than the method or underlying worldview. They thereby abandon the forced dichotomy between constructivism and post-positivism; they also argue that metaphysical concepts such as reality should be disregarded, and lastly that a practical research philosophy guides decisions of methodology. Creswell and Plano Clark (2007) add to this by stating that multiple paradigms may be used within a mixed methods research project, once the researcher makes such choices explicit.

3.2.2 Justification

A need existed in the study for the combination of quantitative and qualitative approaches given that alone, neither approach would be sufficient to reach the required insight required in the current study. The qualitative approach was used to provide detailed information to assist in the interpretation of participant behaviour and to allow for formation of a broader and more contextually relevant understanding of the quantitative data. At the same time, to overcome the main criticisms of a qualitative approach in this context (related to validity, subjectivity, generalisabiliy), quantitative methods were used to provide findings that could be externally validated in terms of statistical testing and data comparison. Creswell and Plano Clark (2007, p. 62) describe the intent of this design as the combination of "the differing strengths and nonoverlapping weaknesses of quantitative methods (large sample size, trends, generalizations) with those of qualitative methods (small sample size, details, in depth)".

In more general terms, the subjective experiences of participants were essential to the study but lacked the objective validity and consistency necessary for scientific research. Further justification of a mixed-methods approach can be found in the need for quantitative results to be explained and validated by qualitative research, especially as the current study represents an interdisciplinary project where the inclusion of the human involvement in machine translation evaluation was an important part of the remit of the overall research project as discussed in Chapter Two.

3.3 Theoretical Framework

3.3.1 Research Questions and Hypotheses

Given the main research question of the study:

- *Does the implementation of linguistic pre-processing in the form of a controlled language rule set result in higher levels of readability and comprehensibility in Statistical Machine Translation output?*

The null hypothesis (H_0) states that: no significant increase in readability or comprehensibility is observed in the controlled condition over the uncontrolled condition in the context of SMT. The alternative hypothesis (H_1) postulates that a significant increase in readability and comprehensibility would be found for the controlled condition over the uncontrolled text in the context of data-driven MT.

From the operationalisation of the concepts of readability and comprehensibility, the following embedded and more specific research questions arise:

- *Does implementation of CL result in improved scores as measured by the traditional readability indices Flesch and LIX?*
- *Are differences in eye tracking measures reported between the uncontrolled and controlled conditions?*
- *Do post-task human evaluation and recall testing show an improvement in readability and comprehensibility after implementation of CL?*
- *Do all of the above measures correlate and yield consistent findings?*
- *What is the relationship between human and machine evaluation of MT in this context?*

3.3.2 Operationalisation

In order to establish objectivity, it is necessary to attempt to operationalise variables in all research projects. Coolican (1996, p. 25) describes how "an operational definition of variable X gives us the set of activities required to measure X". Frey *et al*. (1999, p. 94) describe the activities or observable characteristics as threefold:

- There must be an adequate definition of the characteristics under observation;
- The definition must be valid and accurate;
- The definition must be clear to readers and future users of the research.

In this study, the focus is on the readability and comprehensibility of machine-translated text. As indicated in Chapter Two, the direct measurement of these concepts is not possible, and it can be difficult to separate the concepts of readability and comprehensibility. Therefore, it was necessary to specify the observable characteristics of these concepts prior to their inclusion in the research design. Drawing from the literature and from the lessons learned from the pilot study, the concepts are operationalised as described in Table 3.1 with further detail provided in the later subsections.

Readability	**Comprehensibility**
LIX, and Kandel and Mole's adaptation of Flesch (readability indices)	Recall test
Eye tracking metrics	Post-task questionnaire (5-point Likert scale)
Retrospective questionnaire (5-point Likert scale)	

Table 3.1: Dependent Variables and Respective Measures

3.3.3 Method Design

A biphasic mixed-methods approach was used in this study to examine the effect of controlled language rules on the readability and comprehensibility of machine translated text by means of a participant-centred approach (eye tracking, recall test, Likert scale evaluations) in conjunction with textual metrics namely readability indices and automatic evaluation metrics of MT output. A triangulation mixed-methods design was used in both phases i.e. the pilot study and the main study. This type of design allows different but complementary data types to be collected on the same topic (Creswell and Plano Clark 2007, p. 62). The model employed in this study is best described in Creswell and Plano Clark (ibid., p. 63) as the concurrent model of the triangulation design whereby quantitative and qualitative data are collected concurrently and the two data sets are combined for analysis and interpretation.

3.3.4 Sampling

The current study relied on a self-selected sample of participants and its sampling method therefore falls into the category of opportunity/convenience sampling in that it relied upon suitable participants in the locale to be recruited via correspondence, in this case e-mail. Convenience sampling makes fewer demands on time and money than other sampling methods. The population was envisaged to be users of technical documentation for anti-virus and security software. This documentation was intended for a (European) French speaking audience, and originally written by in-house technical writers in English. Participants were recruited via e-mail through the mailing lists of the Centre for Next Generation Localisation and the Alliance Française in Dublin, and via a French language group mailing list in Dublin.

As evident in the review of literature in the previous chapter, smaller sample sizes are commonplace in eye-tracking studies given that particular skills are required for participation in studies e.g. translation, post-editing, or other linguistic skills. O'Brien (2009, p. 255) reports that the average number of participants in the eye-tracking studies of translation that she had reviewed was twelve. With regard to skill sets for the current study, participants were required to be native speakers of French, have basic computer skills, and not to have any previous experience in the domain of anti-virus or security software. In addition to this, as the current study required participants to evaluate on-screen text in their native language and use a keyboard, a larger sample size was possible as specialised training (e.g. in translation or post-editing) was not required.

The time required to process and analyse eye-tracking data tends to militate against large samples (ibid.). Drawing from the experience of the pilot study, it was estimated that a larger sample size was possible in the time frame. Moreover, there is the risk of poor data capture during eye-tracking experiments, which suggests that researchers should involve more participants in their experiments than they will ultimately report on in their final analyses. O'Brien (ibid., p. 263) assumes "a 30% drop-out rate (approximately) due to a lack of suitability (physical, competence, white coat effect, etc.)". Coolican (1996, p. 43) states that when investigating an experimental independent variable's

effect on most people, a size of twenty-five to thirty is appropriate and that "if significance is not shown then the researcher investigates participant variables and the design of the study". Given that larger samples decrease the likelihood of sampling bias and other associated errors (ibid., p. 42), a sample size of twenty five was believed to be the most appropriate for the main study.

Evidence from the pilot study also supports the above findings. Of twelve participants, data from two participants were discarded due to poor quality. Following the above recommendations, the sample size of the main study was twenty-five with the expectation that a maximum of five participants' data would be discarded thus allowing for twenty participants, which is sufficient for generalisable and valid data analysis. Each participant was randomly assigned into a group (aka condition): uncontrolled condition (n=13) where the uncontrolled output was viewed, and the controlled condition (n=12) where the controlled text was viewed. Sufficient quality data from ten participants was the aim for each condition, and quality assessment resulted in not using data collected from five participants overall (see Chapter Five for a discussion on data quality).

3.3.5 Validity

Data validity is essential in both quantitative and qualitative methods. It encompasses internal and external factors (see Table 3.2) and establishes whether the results of a study meet the requirements of the scientific method. This method requires, among other things, the consistent and accurate function of instruments, which, in this case, was accomplished by means of a review of past usage of instruments in the literature, as well as using established standards in statistical analysis to analyse data gathered in this study. Creswell (2003) proposes triangulation as a means of establishing validity in mixed-method studies, and Frey *et al*. (1991, p. 24) postulate that measurement validity and reliability can both be increased by means of triangulation.

As stated above, validity can be sub-divided into internal and external factors. Internal factors are those related to the experiment design, the researcher's own influence on the experiment and the conclusions drawn from the study (Frey *et al*. 1991, Coolican 1996). External validity pertains to the extent that the findings of the study can be validly generalised (Coolican 1996) and can be discussed under the headings of sampling, ecological validity, and replication (Frey *et al*. 1991). Sampling has already been addressed above. A few points are made here about ecological validity.

Given that the type of text used in the current study is likely to be read on screen, the set-up can be said to be a naturalistic environment for participants. However, due to constraints requiring the eye tracker to remain in the research lab, the study took place in an unobtrusive, quiet, and undisturbed working space, similar to that of a modern office, and therefore may not be a natural environment to all participants, but ensured a consistency for all participants. The computer ran under the Windows XP operating system and the texts were presented on a blank screen, as described in detail below. All precautions taken to minimise threats to validity are summarised in Table 3.2.

Internal Validity	**External Validity**
Experimental setup and environment kept constant; all participants receive the same treatment by the researcher and are exposed to the same scripted experiment protocol (Frey *et al*. 1991)	All methods address the same research question with the same hypotheses
Information clearly presented to participants, who were all fully briefed on what was required of them (Frey *et al*. 1991)	All data are gathered together to form datasets allowing for more accurate interpretation of results
The capture of both quantitative and qualitative data from the same sample in the same session	Use of unobtrusive and ecological means of collecting data (e.g. eye tracking, standard desk and computer layout)
Adequate sample size	Realistic environment with time for familiarisation
Participants unknown to the researcher and having no relationship, personal or professional with him	Information on materials used, conditions, settings, and participants documented in detail to ensure replication (Frey *et al*. 1991)

Table 3.2: Description of Measures to Support Validity

3.3.6 Measurement Validity and Reliability

Measurement validity refers to the extent to which a measurement actually measures the concept (variable or construct) under study (Frey *et al*. 1999, p. 199, Coolican 1996, p. 56). Measurement reliability refers to the consistency of the measurement of a concept (variable or construct). To provide support for both factors, an extensive review of related literature has been carried out to ensure that the measurements of the concepts central to this study are accepted as valid and reliable measures. Frey *et al*. (1999) recommend the use of participant observations, questionnaires, interviews, and pilot testing to ensure reliability by reducing the occurrence of mistakes made by the researcher. As detailed in the next chapter, a pilot study was carried out to this end and brought to light several issues which required consideration before proceeding on to further studies.

3.4 Readability

Drawing upon the literature in Chapter Two, readability is defined here as the extent to which a text can be easily read in terms of linguistic elements (such as number of syllables, number of words and sentences). It is assumed that these elements will influence the reader's interaction with the text, but readability is operationalised as a text-dependent (and reader-independent) attribute.

As already described, the following two indices were used for measuring readability:

- LIX
- Kandel and Moles' adaptation of the Flesch index

Designed as an interlingual measure, the LIX index is computed as follows:

$$A/B + (C \times 100)/A$$

Where A = number of words, B = number of periods, C = number of long words (more than 6 letters) (Björnsson, 1971).

This gives a text a score which falls into one of five categories (or standards): very easy texts (a score of <25), easy texts (25-35), average texts (35-45), difficult texts (45-55), and very difficult texts (>55) (Björnsson 1983). Kandel and Moles' adaptation of the Flesch index (Kandel and Moles 1958) is calculated as follows:

$$209 - (0.68 * (\text{syllables/words})) - (1.15 * (\text{words/sentences}))$$

Both indices were calculated electronically to generate the scores for the respective French texts. As both measures require texts to be greater than 100 words, the paragraphs used in the main study were approximately 150 words long. This also enhanced the ecological validity of the texts as full coherent paragraphs were chosen.

3.5 Comprehensibility

Following Van Slype (1979, p. 62), comprehensibility is defined here as the extent to which a text is understandable. It is classified here as an attribute of the text which is relative to and dependent on the reader, i.e. it can change depending on the reader (reader-dependent) whereas readability as measured above is anchored to the text. This variable is tested by a recall test administered post-task (described below in this section). Figure 3.1 provides a visual description of the conceptualisation of both readability and comprehensibility as operationalised in the current study.

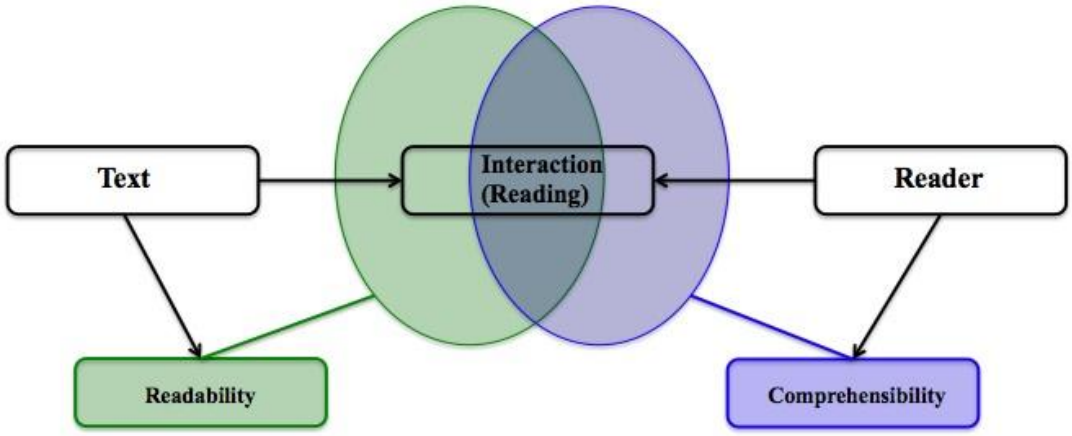

Figure 3.1: Interaction of Attributes of Text and Reader

In creating a questionnaire to test recall, several issues were considered, such as the preference for specific, close-ended, and non-suggestive questions (Converse and Presser 1986). Direction was also taken from authors such as Rapley (2004), who describes how initial questions are generated from the relevant literature, along with the researcher's own beliefs of what areas are of interest in the interview. Similarly, Oppenheim (1966) describes the funnel approach to questioning in which a broad question is asked to begin with, followed by related and more specific questioning.

In the context of the evaluation of MT, Armstong *et al*. (2006) test a range of question types. Bowker and Ehgoetz (2007), Trujillo (1999) and FEMTI

(2010) describe characteristics of questions designed to test comprehension and also describe the use of general questioning followed by more specific questions or internal checks to check for full comprehension or partial comprehension. As described in Chapter Two, recall is divided into three categories: free, cued, and recognition. It was therefore necessary to test each of these levels to ascertain the level of recall. In this light, two general questions were designed for each of the six paragraphs. Each of these (eight) questions tested the three forms of recall discussed in Chapter Two. To begin, a specific question was asked to test cued recall, and then recognition was tested in the form of cloze testing, followed by free (spontaneous) recall via an open-ended question. Figure 3.2 provides an example:

	1. À propos de Symantec AntiVirus		
1.1	After reading this paragraph, do you understand the options for installing Symantec AntiVirus?	Yes	No
1.2	Votre administrateur exécute des analyses sur votre ordinateur et peut configurer des _____ _____.'		
1.3	What are the consequences of an administrator managing the installation?		

Figure 3.2: Excerpt from Recall Test

Following Conway *et al*. (1998, p. 776): the best measurement of recall includes scoring procedures that "exhaust the information collected with a task" and partial-credit scoring. In other words, correct responses to items within a group are assigned a number; all other responses are assigned another with partial recall being rewarded on a proportional basis. Building from this, each general question is worth 1 mark, and intends to test general comprehension. The specific questions are worth 2 marks each, allowing for partial-scoring. The logic of the weighting is that it rewards deeper levels of comprehension. It must also be noted that because each group was exposed to a separate condition, the recall test was modified to contain the content of the respective conditions.

Internal reliability of the recall test was ascertained by means of split-half and Cronbach's α scores with the conventional threshold of 0.7; a value of 0.707 was found. Both methods are single-administration methods (when one test is carried out at one time) of internal consistency and reliability. The split-half method involves randomly choosing half of the sample's scores and comparing

them with the other half - the correlation between the scores shows the reliability. However, this value depends on which scores are chosen for each half. A further step to solve this issue is to calculate every possible split-half reliability by having every possible combination of items to find the average, or in other words the coefficient α. A coefficient $\alpha \geq 0.7$ is "generally accepted as evidence of a satisfactory level of internal consistency" (Howitt and Cramer 2008, p. 408).

3.6 Additional Factors

3.6.1 Task Motivation

The importance of motivation for reading and the reader's interest has been demonstrated by Schallert and Reed (1997) for example. Motivation is of particular relevance to experimental research environments where readers may not have the same motivation or reasons for reading a text as they would in their normal environment. Task motivation is a vital aspect of any experiment concerning humans, especially when mental exertion is concerned. During the briefing stage, participants were informed that they must read for comprehension and that they would be tested afterwards. As participation was self-selected and voluntary, it was assumed that all participants were sufficiently motivated to complete the task as instructed.

3.6.2 Reader Type

As detailed in the previous chapter, people read texts in different ways and have different reading behaviours depending on the situation, e.g. time constraint, purpose of reading. By not imposing a time constraint, the time variable was standardised in that all participants were self-paced. Participants were all given the same instructions and materials were constant across experimental conditions, i.e. those in the uncontrolled condition received the same materials as the controlled condition, except for the output from the respective MT system. Reader type is acknowledged as an additional factor in the study, but it was beyond the scope of the study to obtain a sufficient and representative sample size of participants of different reader types.

3.6.3 Domain Knowledge

As previously mentioned, domain knowledge can compensate for poorly written text and allow a reader to overcome difficulties by accessing their

memory or cognitive schema of the task in question. Participants were recruited on the basis that they had no prior domain knowledge, an advertised prerequisite to the participation in the study.

3.6.4 Time Constraint

As already indicated, because time-limits have been shown to have an impact on task behaviour, in particular, reading behaviour, it was decided that no time limit would be enforced in the study. As evident from the findings of the pilot study, large differences were observed in task time between those with linguistic training and those without. It was, therefore, inadvisable at this time to add an additional factor of time constraints to the experiment, especially one that may have a large influence over the behaviour of the participants, and their thoroughness and success of their tasks. Finally, the duration of the task is largely dependent on each individual participant.

3.6.5 Word Frequency

While word frequency (which captures whether or not the words used in a text are common in the language in general) is a factor in some proposed measures of readability, it is not used in the study as the two conditions in the current study are represented by texts that differed only in the edits made to resolve the CL violations (see section 3.8.3 below), and no significant difference could be found between the two conditions in terms of word frequency. Additionally, as the corpora came from the anti-virus/technical support domain, the comparison of the MT output created from these corpora against bands of word frequencies obtained from a general-language corpus may not have been valid. To the knowledge of the researcher, no validated corpus is available for the technical support or other comparable domain in the French language. Lastly, in a study of text complexity using readability indices, Jensen (2009) found that word frequency correlated with both Flesch and LIX scores, and discusses further limitations of using word frequencies.

3.7 Corpus Description

Two corpora were used in this study, one controlled and one uncontrolled corpus. The uncontrolled corpus contained 356,380 words and the controlled corpus 475,375. The corpora consisted of technical support documentation from the domains of anti-virus and security software for two different product lines: SAV 2006 (uncontrolled corpus) and SEP11 2007 (controlled corpus).

The uncontrolled corpus was authored by in-house technical writers at Symantec in English and translated by in-house translators into several languages. It represents the company's general style rules (see Appendix E) and is of the same domain as the controlled corpus but for a different product and was written before the implementation of CL rules. The controlled corpus was also authored by in-house technical writers in English, who used controlled language rules and CL checker *acrocheck* (version 3.1) to ensure compliance to said rules (see section 3.8.1).

The uncontrolled corpus was translated into French by in-house human translators and the controlled corpus was translated into French by Symantec's customised RBMT system Systran (version 5.0). This output was post-edited by in-house human post-editors to a publishable standard. Both corpora went through similar quality assessment prior to dissemination. At this point it is acknowledged that the use of post-edited MT output is not ideal for training an SMT system; however, no human translations of a controlled corpus were available as the supplier, like most industrial vendors, implemented the use of controlled language alongside machine translation.

Kennedy (1998) highlights the importance of corpus design and compilation in the validity and reliability of associated research. He (ibid.) pinpoints three issues of importance: the permanence of the corpus (if it is static or dynamic); the representativeness of the corpus (how accurately does the corpus represent a particular language, genre etc.); and the corpus size. The corpora used in this study are static, and consist of product information and guidelines from two product lines and so accurately represent real-life usage of language in this domain.

Using Wordsmith Tools (version 5.0) the following information was gathered: 'tokens' in Table 3.3 gives the number of orthographic words contained in the corpus; while 'types' refers to the number of different words. The type-token ratio is the relationship between the total number of words (tokens) and the number of these words that are different (types). Wordsmith Tools provides a standardised ratio whereby the ratio is calculated per 1,000 words and then an overall average is given. Word/sentence length measures the number of words/sentences and their average length. Lexical density is the proportion of the tokens in the corpus accounted for by content (or lexical) words, as opposed to function (or grammatical) words (Table 3.3).

	Uncontrolled	**Controlled**
Tokens	475,375	356,380
Types	8,612	5,730
Standardised Type/Token Ratio	1.82	1.68
Mean Word Length (in characters)	5.21	5.28
Sentences	485	319
Mean Sentence Length (in words)	19.48	16.19
Content Words	304,866	245,774
Function Words	170,509	110,606
Lexical Density	64%	69%

Table 3.3: Corpora Metadata

Following Bowker and Pearson (2002) the corpora can be further characterised based on the data in Table 3.4:

Information	**Uncontrolled**	**Controlled**
Text Extract Vs Full Text	Full Text	Full Text
Medium	Written, XML format	Written, XML format
Subject	Security	Security
Text Type	Anti-virus and security technical support documentation	Anti-virus and security technical support documentation
Authorship	In-House Technical Authors	In-House Technical Authors
Translated	In-House	In-House
Languages	English-French	English-French
Publication Date	2006	2007

Table 3.4 Additional Corpora Information

Lastly, a word frequency list was generated from each corpus – Table 3.5 highlights the 25 most frequent words with function words removed. The # symbol refers to numbers that appear in the corpus that have been disregarded

by the software as it has been set to ignore individual sequences of numbers, e.g. version 2.5 and version 2.6 are both classified under #. Figure 3.3 illustrates the frequency of words containing different numbers of letters (from 1 letter per word or LPW to 50 LPW).

Rank	Uncontrolled	Controlled
1	#	#
2	SYMANTEC	IS
3	CLIENT	CLIENT
4	SERVER	CLICK
5	IS	SYSTEM
6	SECURITY	SERVER
7	FILE	ARE
8	FIREWALL	POLICY
9	ARE	SYMANTEC
10	FILES	NOT
11	CLICK	NETWORK
12	OPTIONS	USE
13	SYSTEM	SELECT
14	ANTIVIRUS	PROTECTION
15	SETTINGS	SETTINGS
16	COMPUTER	SECTION
17	NOT	C
18	CLIENTS	COMPUTER
19	SCAN	FILE
20	GROUP	DOCBOOK
21	VIRUS	DOCBOOKX
22	SERVERS	DOCTYPE
23	NETWORK	DTD
24	INFORMATION	ENTITYDECLARATIONS
25	USING	LOG

Table 3.5: Corpora Word Frequency

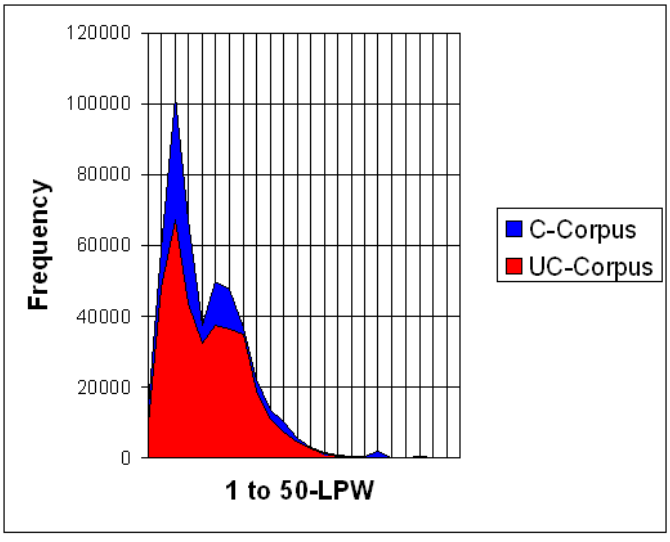

Figure 3.3: Frequency of Words from 1 LPW to 50 LPW

3.8 Controlled Language

As stated in the previous chapter, the primary intent of a CL is the improvement of text quality by enforcing "constraints on lexicon, grammar, and style" (Huijsen 1998, p. 2). Additional and more specific aims, however, will obviously differ depending on the intended application. The corpora and CL rule set used in this study have been donated by the software security company, Symantec. Symantec uses a style guide and, more recently, a set of CL rules (see below), which contain both human and machine oriented controlled language rules. Gains made from the implementation of this setup have been reported by Roturier (2009), who demonstrates the benefits of using CL in an industrial MT environment in terms of time and cost reduction as well as output improvement and faster post-editing.

3.8.1 Controlled Language Checker

As indicated in Chapter Two, a CL checker is a software application designed to highlight linguistic structures that do not comply with a predefined list of rules. The CL checker used to author the source texts used in this study was a product called *acrocheck* (developed by *acrolinx IQ*, see http://www.acrolinx.com), a package that has been customised for use by Symantec. Authors use this program to enforce Symantec's CL rules. Prior to the implementation of these rules, Symantec used a style guide, the contents of which are also listed later in this section, and which subsequently became part of the rule set currently in use. The authors of the controlled corpora in this project used the *acrocheck* tool to ensure conformity to the rules below. This can be seen as a pre-processing stage in the MT process, which, it can be hypothesised, will affect the quality and therefore the user-based reception of the output.

3.8.2 Controlled Language Rules

The following are some examples from a list of authoring rules contained in Symantec's style guide and which have been used in the authoring of the controlled corpus used in this study. They can be subdivided into subcategories: spelling, grammar, style, and MT. It was not possible to obtain a detailed report of which individual rules were applied in the creation of the corpora. A full list of all possible rules and their categories is provided in Appendix E.

- Distinguish between "a" and "an"
- Use one space after sentence end
- Avoid passive voice
- Avoid progressive tense
- Use articles
- Write positive statements
- Use "could" and "if" in conditional clauses
- Avoid slashes
- Avoid parenthetical expressions

The rules were developed by the CL vendor (*acrolinx IQ*) in partnership with the user (in this case Symantec) and vary from general rules to MT-specific rules that have been customised specifically for the RBMT system in use on-site, i.e. Systran.

3.8.3 Application of Controlled Language Rule Set

To account for the results to follow in this chapter, the differences between the uncontrolled and controlled texts are first detailed. As previously described, the *acrocheck* tool was used to check the uncontrolled source text for violations of the Symantec rule set. An overview of the violations shows no mistakes in spelling or grammar, but 33 violations of the style subsection of the rule set – see Table 3.6, which shows a summary of rule violations found by the

researcher using the *acrocheck* tool. The tool also provides its own gauge on quality using a score based on the number of occurrences of each violation and their perceived severity. Unfortunately, the measuring and logic behind this process are proprietary and unavailable to the researcher. During the course of using the tool many inconsistencies were found, e.g. where one rule is applied to several instances but not others. This may highlight problems in the tool's own measurement of quality, however, without access to the above information further investigation was curtailed – the results of the tool are nevertheless of relevance and interest here. The information provided states that a score of < 100 gives a green flag, 101 to 199 results in a yellow, and anything above 200 warrants a red flag. In this case, with a score of 301, the uncontrolled text would require editing before it could be deemed to be of sufficient quality to proceed further in the workflow, e.g. for MT or human translation.

Rule	Frequency
Avoid Passive	4
Sentence Too Long	12
Disambiguate 'ing' Words	8
Use Articles	3
Avoid Slashes	1
Use Relative Pronoun	1
Avoid Unnecessary Words	1
Avoid Future Tense	1
Avoid 'Could'	1
Avoid Sentence Beginning with "Or"	1
Total	33

Table 3.6: Summary of Violations in the Uncontrolled Source Text

Table 3.7 shows each violation of the respective rule, its location by paragraph number, and its uncontrolled vis-à-vis its edited controlled version. The bolded and underlined text represents the violations highlighted in the *acrocheck* tool. With the exception of the 33 sentences, both texts were identical. In some cases, resolving a rule violation resulted in further violations (see, for example, row 23 in Table 3.8). In such instances the violation equating to the lower negative score by *acrocheck* was given preference, i.e. aiming for a better score overall. The *acrocheck* can use the term sentence in the rather vague sense of segment.

Row	Rule Violation	Paragraph	Uncontrolled	Controlled
1	Avoid Passive	1	A stand-alone installation means that your Symantec AntiVirus software **is** not **managed by** a network administrator.	A stand-alone installation means that a network administrator does not manage your Symantec AntiVirus.
2	Sentence Too Long	1	A stand-alone computer that is not connected to a network, such as a home computer or a laptop stand-alone, with a Symantec AntiVirus installation that uses either the default option settings or administrator-preset options settings	A stand-alone computer that is not connected to a network with a Symantec AntiVirus installation that uses either the default or administrator-preset options settings
3	Disambiguate 'ing' Words	1	A remote computer that connects to your corporate network that must meet security requirements before **connecting**.	A remote computer that connects to your corporate network that must meet security requirements before it connects
4	Use Articles	1	However, you may want to adjust them to suit your company's needs, to optimize system performance, and to disable **options** that do not apply.	However, you may want to adjust them to suit your company's needs, to optimize system performance, and to disable the options that do not apply.
5	Avoid Passive	1	If your installation **is managed by** your administrator, some options may be locked or unavailable, or may not appear at all, depending upon your administrator's security policy.	If your administrator manages your installation, some options may be locked or unavailable, or may not appear at all, depending upon your administrator's security policy.
6	Sentence Too Long, Avoid Slashes	2	The Technical Support group's primary role is to respond to specific questions on product **feature/function**, installation, and configuration, as well as to author content for our Web-accessible Knowledge Base.	The Technical Support group's primary role is to respond to questions on products and to author content for our Web-accessible Knowledge Base.
7	Sentence Too Long	2	For example, the Technical Support group works with Product Engineering as well as Symantec Security Response to provide Alerting Services and virus definitions updates for virus outbreaks and security alerts.	For example, the Technical Support group works with other groups to provide Alerting Services and virus definitions updates for virus outbreaks and security alerts.
8	Sentence Too Long, Use Relative Pronoun	2	Global support from Symantec Security Response experts, which is available 24 hours a day, 7 days a week worldwide in a variety of languages for **customers enrolled** in the Platinum Support Program	Global support from Symantec Security Response, 24 hours a day, 7 days a week in a variety of languages for those enrolled in the Platinum Support Program.
9	Disambiguate 'ing' Words, Avoid Unnecessary Words	3	When you install and run a Trojan horse, it appears to be **performing** a helpful function, but it is **actually damaging** your computer's operating system.	When you install and run a Trojan horse, it appears to perform a helpful function, but damages your computer's operating system.
10	Disambiguate 'ing' Words	3	Default Trojan horse rules are always **blocking** rules, in contrast to General or Program rules, which may permit access.	Default Trojan horse rules always block, in contrast to General or Program rules, which may permit access.
11	Use Articles	3	Trojan horse rules work by matching **attack patterns** associated with a list of known threats against ongoing network	Trojan horse rules work by matching the attack patterns associated with a list of known threats against ongoing

Row	Rule Violation	Paragraph	Uncontrolled	Controlled
			communications.	network communications.
12	Sentence Too Long	3	Occasionally, harmless network activity can trigger a Trojan horse alert, if the communication involves using specific ports or other criteria associated with a known Trojan horse.	Occasionally, harmless network activity can trigger an alert, if the communication involves using specific ports or other criteria associated with a known Trojan horse.
13	Sentence Too Long, Avoid Passive	3	If you continually receive the same Trojan horse alert, you may want to investigate further to make sure the alert is not being **generated** by normal activity or communications on your network.	If you continually receive the same alert, you may want to ensure that normal activity or communications on your network does not generate the alert.
14	Avoid Passive	4	Delete files that are **infected** by viruses in the Quarantine	Delete infected files via Quarantine
15	Sentence Too Long	4	Deleting a file that is infected by a virus reduces the threat that a virus might spread by removing the file (and thus the virus) from your computer.	Deleting an infected file reduces the threat that a virus might spread by removing the file and virus from your computer.
16	Sentence Too Long	4	Because viruses can damage parts of a file, deleting the infected file and replacing it with a clean backup file may be better than cleaning the infected file.	Because viruses can damage parts of a file, deleting and replacing it with a clean backup file may be better than cleaning the infected file.
17	Disambiguate 'ing' Words	5	**Enabling** and **disabling** Auto-Protect	To enable and disable Auto-Protect
18	Sentence Too Long	5	It checks programs for viruses and security risks as they run and monitors your computer for any activity that might indicate the presence of a virus or security risk.	It checks running programs for viruses and security risks and monitors your computer for any suspicious activity.
19	Avoid 'Could'	5	When a virus, virus-like activity (an event that **could** be the work of a virus), or security risk is detected, Auto-Protect alerts you.	When a virus, virus-like activity (an event that may be the work of a virus), or security risk is detected, Auto-Protect alerts you.
20	Disambiguate 'ing' Words	5	For example, this might occur when you are **installing** new computer programs.	For example, this warning might occur when you install new computer programs.
21	Avoid Future Tense, Disambiguate 'ing' Words	5	If you **will** be **performing** such an activity and want to avoid the warning, you can temporarily disable Auto-Protect.	If you perform such an activity and want to avoid the warning, you can temporarily disable Auto-Protect.
22	Sentence Too Long	6	Your administrator might lock Auto-Protect so that you cannot disable it for any reason, or specify that File Auto-Protect can be disabled temporarily, but reenables automatically after a specified amount of time.	Your administrator might lock Auto-Protect so that you cannot disable it, or specify that it can be disabled temporarily, but reenables automatically after a specified time.
23	Sentence Too Long	6	Inclusions and exclusions help you to balance the amount of protection that your network requires with the amount of time and resources that are required to provide that protection.	Including and excluding objects can help you to balance the amount of required protection with the amount of resources necessary to provide that protection.
24	Sentence Too Long	6	For example, if you choose to scan all file types, you might want to exclude certain folders that contain only data files that are not subject to viruses.	E.g. if you choose to scan all file types, you might want to exclude folders containing the data files that are not subject to viruses.

Row	Rule Violation	Paragraph	Uncontrolled	Controlled
25	Avoid Sentence Beginning With 'Or', Use Articles	6	**Or**, you might want to scan only the files with **extensions** that are likely to contain a virus or other risk.	Otherwise, you might want to scan only the files with the extensions that are likely to contain a virus or other risk.

Table 3.7: Description of Violations and Edits Made

3.9 Machine Translation

Building upon the discussion of machine translation in Chapter Two, this section describes the system used in the current research, i.e. the MaTrEx system, a hybrid data-driven MT system developed at Dublin City University (see Du *et al*. 2009, Morrissey 2008). As previously stated, compared to other approaches such as rule-based MT, data-driven MT generally favours probabilistic models built from large amounts of bilingual parallel corpora. The corpora used in the current study were discussed in the previous section, and the following will focus on the MaTrEx system in detail.

3.9.1 System Description

The MaTrEx system is a hybrid of statistical MT and example-based MT components (Groves 2007). The underpinning approach of the MaTrEx system is the marker hypothesis which states that "all natural languages have a closed set of specific words or morphemes which appear in a limited set of grammatical contexts and which signal that context" (Green 1979, p. 483). In the context of the MaTrEx system this means that the structure of a natural language can be marked at a surface level and subsequently be deconstructed into smaller chunks which can then be recombined as needed to form different segments from the original. Flanagan (2009) provides a graphical overview of the MaTrEx system (Figure 3.4), which, as outlined by Stroppa and Way (2006), is comprised of four critical components:

- The word alignment module, which is fed a segment-aligned corpus as input and outputs a set of word alignments;
- The chunking module, which takes as input a segment-aligned corpus and outputs source and target chunks;
- The chunk alignment module, which is given the source and target chunks and aligns them on a segment-by-segment level;

- Lastly, the decoder, which searches for a translation to a new input using the original aligned corpus and derived chunk and word alignments.

The hybrid approach of the system is described in Groves and Way (2006) who postulate that the combination of SMT and EBMT approaches realised in one system can improve performance and the quality of the output. By means of automatic metrics, they (ibid.) demonstrate the success of this approach.

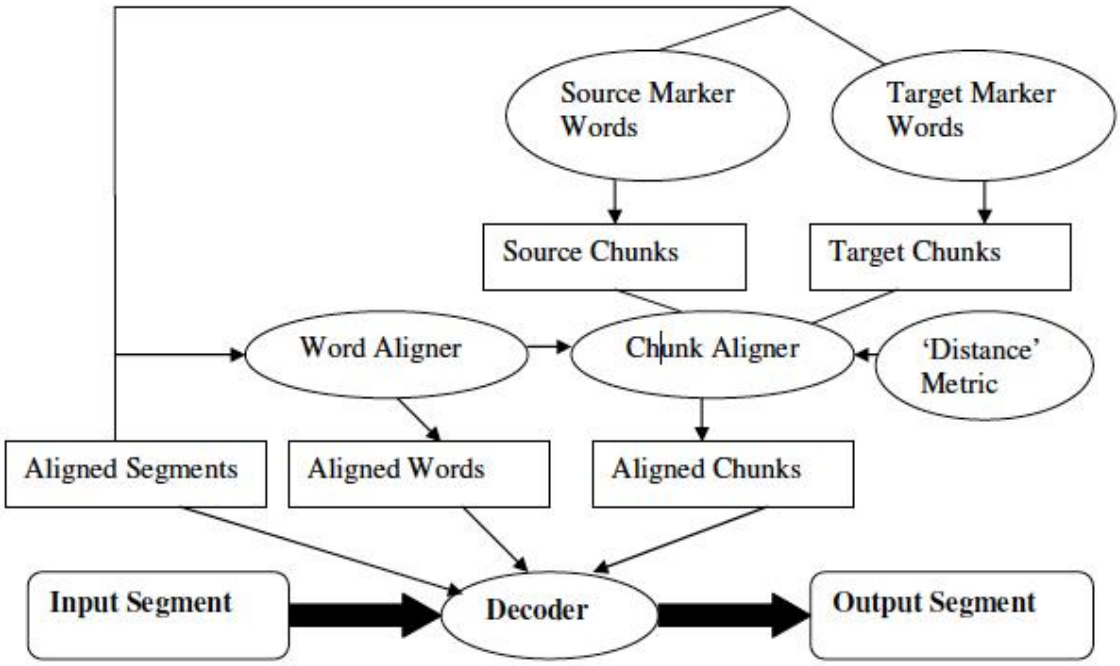

Figure 3.4: The MaTrEx System (Flanagan 2009)

The corpora, as detailed in the following section, were received in .xml and .tmx format and their input into the MaTrEx system required conversion into a simpler text format (.txt). A test set was extracted from the uncontrolled corpus. It consisted of six paragraphs (amounting to 1, 207 words, approximately 0.0025% of the uncontrolled corpus) chosen because:

1. They were between 100 to 200 words long to satisfy the requirements of the readability indices;
2. They could stand alone as independent paragraphs;

3. They were introductory in natures, i.e. they did not contain very technical vocabulary or other items such as command lines;
4. They contained the greatest number of controlled language violations (which would subsequently allow maximal difference between these paragraphs and a controlled version of the same – see below).

The researcher applied the CL rules using the *acrocheck* (version 3.1) tool to these paragraphs in order to create a 'controlled' version. This design allowed participants in the controlled-language and uncontrolled-language conditions to view translated texts whose source texts differed only by the fact that the controlled language rules had been applied or not.

The remaining 99.0075% of the uncontrolled corpus and the entire controlled corpus was then used to train two unique iterations of the same MaTrEx system, i.e. one 'controlled' system and one 'uncontrolled' system. Although these training corpora were relatively small by SMT standards, they were of high quality as they had been subjected to quality assessment at Symantec and their size is nonetheless comparable to the training corpora reported on in (Du *et al*. 2009) who used approximately 500,000 words to train the MaTrEx system for the English-French language pair and performed well in the shared task competition at the European Association for Computational Linguistics (EACL) 2009.

3.9.2 Automatic Evaluation Metrics (AEMs)

In the current study, three AEMs were used throughout: GTM (General Text Matcher) (Turian *et al*. 2003), BLEU (BiLingual Evaluation Understudy) (Papineni *et al.* 2002), and TER (Translation Edit Rate) (Snover *et al.* 2006). These AEMs were chosen as they are widely applicable for European languages, they appear frequently in the relevant research literature in studies that compare systems, and they are relatively easy and cost-effective to use. Finally, Symantec's own evaluation tool facilitated simultaneous calculation of these metrics.

There are other AEMs that may have been used, such as Meteor (Banerjee and Lavie 2005) and TERp (Snover *et al*. 2009) etc. Additional resources are required for the accurate use of these AEMs, however, such as a database of synonyms e.g. WordNet. Such metrics were not chosen, but may be of interest in future work or later retrospective comparisons.

3.10 Eye Tracking

As evident from the review of the literature, eye tracking has become an established means of measuring cognitive effort for various tasks. While a growing number of studies are adopting eye tracking in relation to translation evaluation, the established body of research in other closely related areas such as translation process studies (e.g. O'Brien 2006, 2008, Göpferich *et al*. 2008), audio-visual translation (e.g. Caffrey 2009, Perego and Ghia 2011) and reading studies (e.g. Rayner 1998), in particular, provides sufficient justification for the use of eye tracking in this study. Eye tracking offers an interesting and completely novel method of evaluating MT output as it enables the cognitive effort involved in reading the target text to be measured in an objective way. Cognitive demands and MT evaluation are largely overlooked in MT research (cf. Roturier 2004); it is generally simply assumed that when a human evaluates MT output as "good", that output is easily read and understood by the end user. The eye-tracking method also offers additional advantages in that the evaluator does not have to be bi-lingual and requires no training in evaluation techniques or criteria and this opens up the possibilities of including real end users in MT system evaluation (see section 3.3.4 for information pertaining to the participants in this study).

3.10.1 Hardware and Software

Fundamentally, an eye tracker monitors and records eye movement and pupillometric information by means of inbuilt infrared diodes that bounce light off the eyes. The Tobii 1750 device was used in both the pilot and main study. It is a non-invasive eye tracker, i.e. participants do not have to wear head-mounted equipment or use head rests or bite bars (used to compensate for head movement during recording). The device uses corneal reflection and infrared diodes integrated within a 17-inch TFT monitor. These diodes bounce light off the eye and thus extrapolate the location of the eye on screen in terms of an X-Y axis; they also allow for measurement of pupil dilation. While the non-invasive nature is usually not intimidating to participants and, presumably, allows them

to behave more normally, the lack of control leads to some level of inaccuracy in the data. To compensate for this a retrospective think-aloud protocol was used to supplement the eye tracking data with additional qualitative data as outlined below.

No universal standard exists for the filtering of the data collected by the eye tracker, although the settings used by the Eye-to-IT (see http://cogs.nbu.bg/eye-to-it) project were adopted in this study as that there would be a common 'yardstick' that would allow comparisons between the current research and at least one existing project whose researchers appear frequently in the literature. Rayner and Sereno (1994) find that average fixations in reading tasks tend to range between 200 and 250ms and suggest a minimal threshold of 175ms to recognise a fixation below which errors would occur. This threshold of 175ms has been used in other translation process studies such as Alves *et al*. (2009) and Jensen *et al*. (2009). However, Alves *et al*. (2009) highlight the issue of filter configuration acting as an intervening variable across eye-tracking studies in translation studies and conclude that further work is required to determine the most accurate settings.

The setting used in the pilot study was 100ms as it was found that, at higher thresholds, fixations on some sentences were not being counted, suggesting that participants did not read the sentence. However, upon changing the setting to the lower value of 100ms, fixations were identified which supported evidence from the pilot study whereby participants commented on their eye-tracking behaviour and had obviously read the sentence. A similar modification was also made by Hvelplund (2011) who encountered a similar issue.

In addition to these settings, general guidelines for the presentation of on-screen stimuli are described in Gerganov (2007), who prescribes a screen resolution of 1280 by 1024 pixels, a font size of at least 20 (on a 17 inch monitor), a font style of Tahoma, double line spacing, and a maximum of 90 characters per line. These guidelines were adhered to throughout the pilot and main studies.

After O'Brien (2009), data from participants who spent more than 70% of the task duration looking away from the screen were discarded from the pilot

study. Further details about assessing the quality of eye tracking data for the main study can be found in Chapter Five.

The software used in the main study was Tobii Studio version 2.2.8, a suite of tools which facilitates the creation, employment, and analysis of eye-tracking projects. This is a newer software package provided by Tobii, the manufacturer of the eye tracker, and was used instead of the Clearview package used in the pilot study.

3.10.2 Metrics

Gaze Time (Observation Length)

Gaze time is the period of time a participant spends gazing within an Area of Interest (henceforth AOI). For this study, AOIs consisted of all data within a 5cm radius of each sentence in order to allow for all possible data relating to the sentence to be captured and to exclude unwanted data, e.g. when participants look at the toolbar or clock. This radius accounts for peripheral vision at the recommended distance from the screen: 60-65cm and ensures all possible data relating to each individual letter are taken into account. The term gaze time is used in the description of the pilot study as the software for the study used this term throughout, e.g. for tables and figures. For the main study, the updated software used the tern observation length; this change has also been adopted here for the description of the main study.

Fixation Count

As already described, fixations are defined as "eye movements which stabilize the retina over a stationary object of interest" (Duchowski 2003, p. 43) and occur when the eye focuses on a particular area item e.g. a word on the screen.

Average Fixation Duration (Length)

The average duration, usually given in milliseconds, of fixations as described above. The terms fixation duration and fixation length are synonymous.

Percentage Change in Pupil Dilation

The change in the size of the pupil measured in units of percentage where the baseline is the average pupil size across the task. The pupil can dilate (become larger) or contract (become smaller) depending upon many factors such as external stimuli, the individual's state, and physiological changes e.g. caffeine.

Regressions

A regression is defined here as "any eye movement that begins at the right-most point the reader has fixated and leaves the currently fixated region to the left" (Pickering and Traxler 1998, p. 945). The number of regressions was counted per sentence and, as using this value alone would yield a limited analysis; the distance travelled for each regression in units of words was attached to each regression to give a *regression distance* value which is used in the analysis. Therefore, the greater the distance travelled, the higher the value of the regression distance. For example, Participant 1 read sentence 2 and a regression was shown from the final word to the previous word. This was followed by a normal linear continuation to finish the sentence. This differs greatly from sentence 3, where the participant's regression led to the rereading of the entire sentence; and such a difference needs to be taken into account. Regressions were counted manually using the Gaze Plot function of Tobii Studio using the following criteria:

- a regression where the gaze left the AOI was not counted;
- a regression to the same word was not counted, as this is a second fixation on the same word, and is captured in the fixation count and length measures;
- where a regression left the sentence and returned to a previous sentence, the sentence in which the following fixation occurred was counted as the recipient of the regression;
- multiple regressions in succession were counted individually, e.g. regressing from word 3 to word 2, then from word 2 to word 1 equates to two regressions each with a regression distance of one.

3.11 Chapter Summary

This chapter presented the methodology used in both the pilot and main studies. Firstly, the underlying philosophical approaches of the study were outlined and justified. This was followed by a description of the theoretical framework which included the research questions and hypotheses of the main study, and the operationalsation of the variables of readability and comprehensibility. It also discussed methods and sampling used in both studies, and overall issues of validity and reliability. Thirdly, the concepts of readability and comprehensibility were described as they are operationalised in this study. This was followed by a discussion of known additional factors that would have impacted the study, i.e. reader type, motivation, domain knowledge, and time constraints. Fourthly, the corpora and controlled language rule set used in the study were described in detail. This was followed by a description of the study's MT system and automatic evaluation metrics used in the study. Lastly, the eye tracking element was explained with descriptions of hardware and software, and definitions for the metrics employed throughout both studies. In order validate this methodology, a pilot study was conducted (Doherty and O'Brien 2009, Doherty, O'Brien, and Carl 2010), the findings of which are discussed in the next chapter.

Chapter Four

Pilot Study

4.1 Chapter Overview

This chapter describes the pilot study, which was designed with the intention of validating the methodology prior to the main study. Firstly, the research questions and hypotheses of the pilot study are outlined as they differ slightly from those of the main study. Details of the pilot study's methodology are then presented, followed by the results of the study and a discussion of same. Finally, overall conclusions with recommendations for the next steps of the project are presented.

4.2 Aims

As stated, the main aim of the pilot study was to test the assumptions, design, technical feasibility, and validity of the methodology before the main study. The controlled language variable was not present in this study, so as to first establish and validate the eye tracking method for use in the evaluation of MT, a hitherto unexplored avenue. The focus of the study was on the use of the eye tracking metrics listed in Chapter Two, namely: gaze time, fixation count, fixation duration, and pupil dilation. It should be noted here that different corpora and MT systems were used from those in the main study, and that participants of the pilot carried out an evaluation in which the operationalisation of readability and comprehensibility as employed in the main study were not used. However, all of these differences are described in detail in the following sections.

4.3 Experiment Design & Methods

A human evaluation was conducted of rule-based MT output from English to French in a previous study on CL and the acceptability of MT output (Roturier, 2006). In this evaluation, four human evaluators were asked to rate output on a scale of 1-4 where 4 signified "Excellent MT Output", 3 signified "Good", 2 "Medium" and 1 "Poor". A full description of the evaluation criteria for that study is available in Roturier (2006). Roturier's (ibid.) corpus contained sentences from the domain of documentation describing anti-virus software. For the purposes of the pilot study reported on here, 25 of the lowest rated (also called 'bad') and 25 of the best-rated (also called 'good') sentences, were selected from Roturier's corpus and randomised to create a new 50-sentence corpus.

The number of sentences was deliberately small since the main goal was to test eye tracking as an MT evaluation methodology and not to actually evaluate the quality of MT output. The underlying hypothesis was that the highest rated sentences would be easier to read than the lowest rated ones. Likewise, it was assumed that the ease with which sentences could be read and understood influenced the scores given previously by the human evaluators, even though they were not asked to evaluate for readability or comprehensibility.

Eleven native speakers of French participated in the study (twelve were recruited and one was eliminated due to poor quality data). All participants were enrolled at the time of the study as full-time students at Dublin City University, some on translation programmes and others on business and computer science programmes. They were not experts, nor did they indicate particular knowledge of the domain of the corpus used in the study. It was assumed that participants would have to exert cognitive effort to construct an internal representation of the meaning of each sentence and that the effort to do so would be higher for the 'bad' sentences and this would, in turn, be reflected in the data recorded via the eye tracker. It was also assumed that the participants had more or less equal reading ability; generally speaking, reading ability, reader type, and prior knowledge may all influence reading behaviour (Daneman and Carpenter 1980, Kaakinen *et al.* 2003), but it was beyond the scope of the pilot study to measure

the effects each of these variables might have on reading MT output or indeed reading in general. However, these factors and their possible effects are discussed in the context of the study towards the end of this chapter.

Prior to commencement, each participant was informed about the study and what would be required. Each participant was allowed time to ask questions and to sign the university's standard Informed Consent Form (found in Appendix A). The participants were first given a warm-up task in which they read five sentences on screen, to allow them to accustom themselves to the environment and avoid initial disturbances (relative to the task) in behaviour and consequently eye-tracking data. They were then presented with the test sentences in a random order so as to avoid longitudinal effects such as fatigue and sequence recognition (i.e. 'bad' and 'good' sentences were mixed, but presented in the same order for all participants) and participants were not aware that sentences had already been rated in a prior human evaluation task. They were asked to read the sentences for comprehension and, since motivation is an important factor in reading (Kaakinen *et al*. 2003), were informed that they would be asked some questions at the end to see if they had understood the sentences. The sentences were presented using a tool called Translog. Translog was originally developed for researching human translation processes (Jakobsen 1999), but has recently been modified to interface with an eye-tracker and other tools developed within the EU-funded Eye-to-IT project. The Translog tool allows text to be displayed in a window in a similar fashion to a text editor. The participants pressed the "Return" key when they wanted to move to the next sentence and no time pressure was applied as this has also been shown to have an impact on reading behaviour.

The sentences were read in isolation for two main reasons: (i) it is easier to measure fixation count, duration, pupil dilation etc. when only one sentence appears on the screen at any one time. This allowed us to increase measurement validity, but obviously reduced ecological validity since readers normally read "text" rather than isolated sentences; (ii) this scenario reflected the initial human evaluation in Roturier (2006) where individual sentences (and not whole texts) were evaluated. As the focus here was on fluency, only the MT output in the target language (i.e. French) was presented and not the reference translation,

which is commonly presented in human evaluation of translation and indeed machine translation. Consequently, participants were not asked to evaluate adequacy in this study. This method allowed for evaluation by monolingual MT users, monolingual evaluations are likely to be useful in scenarios where, for example, end-users of technical support documentation have to express their satisfaction or dissatisfaction with MT output.

As described in the previous chapter, the Tobii 1750 eye tracker was used in conjunction with Translog. While the non-invasive nature of the Tobii 1750 increases the validity of the online reading experience, the lack of control leads to some level of inaccuracy in the data. An attempt was made to compensate for this by also using retrospective think-aloud protocols. Experimental conditions such as distance from monitor, temperature, noise, and lighting were kept constant. The analysis software used to analyse the eye tracking data was ClearView (version 2.6.3), which also produces an AVI (video file) of the reading session that displays the eye movements and fixations for each participant overlaid on the text. This was played back to the participants immediately after the session in Camtasia Studio (screen recording software) and they were asked to comment on their reading behaviour. This commentary was recorded and all utterances were transcribed and coded as either: 'All Positive', 'All Negative', 'Mixed', 'Silence', or 'N/A', where mixed comments contained both positive and negative content, and N/A denote where no comments were made. These categories were chosen for their relevance to the types of responses given by participants. Categorisations could then be compared with the ratings of the original evaluators and automatic evaluation metrics; this comparison required conversion of categorisation into numerical values for statistical data analysis.

The formal hypotheses of the pilot study were that the quality of the MT output would be reflected in the eye tracking data. More specifically:

- Gaze time would be longer for sentences rated as 'bad' quality MT output;
- Average fixation count would be higher for sentences rated as 'bad';
- Average fixation duration would be longer for sentences rated as 'bad';
- Pupil dilation would be larger for sentences rated as 'bad';
- Participants would agree with the original human evaluation.

4.4 Results

The results from this pilot study are presented in Doherty and O'Brien (2009) and Doherty, O'Brien, and Carl (2010). Overall, a reasonable level of support was found for the hypothesis that MT quality would be reflected in the eye tracking metrics. However, fixation duration and pupil dilation did not correlate to a significant degree with MT quality and therefore the research design required further consideration prior to the main study, as detailed in the next section. As indicated in Chapter Two and Three, gaze time is the period of time a participant spends gazing within an Area of Interest (AOI). For this study, the AOIs were defined around each sentence in order to capture all possible data relating to the reading of the sentence. The total gaze time per participant, given in minutes, is presented in Figure 4.1. The average was 5.23 minutes (median = 5.06):

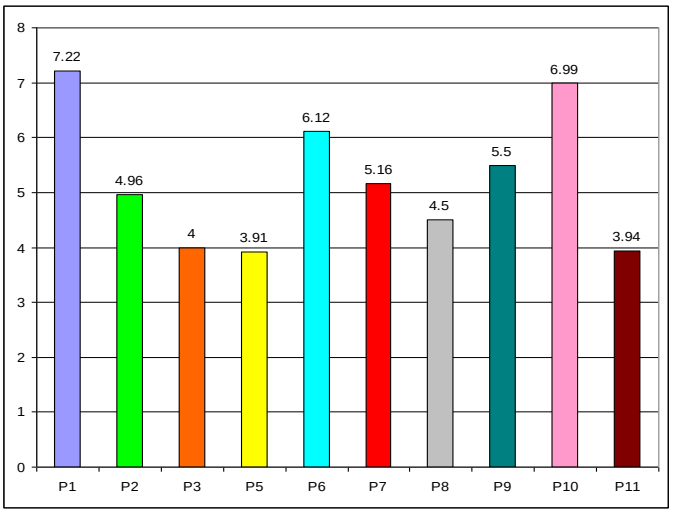

Figure 4.1: Total Gaze Time for All Participants (in minutes)

Figure 4.2 shows the average gaze time per sentence across all participants in milliseconds. As hypothesised, the 'bad' sentences had longer gaze times than the 'good' sentences.[6]

[6] The following figures are of box plots (or box-and-whisker plots) which graphically represent numerical data via five line summaries which represent (from bottom to top): the smallest observation (or sample/range minimum), the lower quartile (Q1), the median (Q2), the upper quartile (Q3), and the largest observation (or sample/range maximum). In the description, the x-axis variable is described firstly, followed by the y-axis variable.

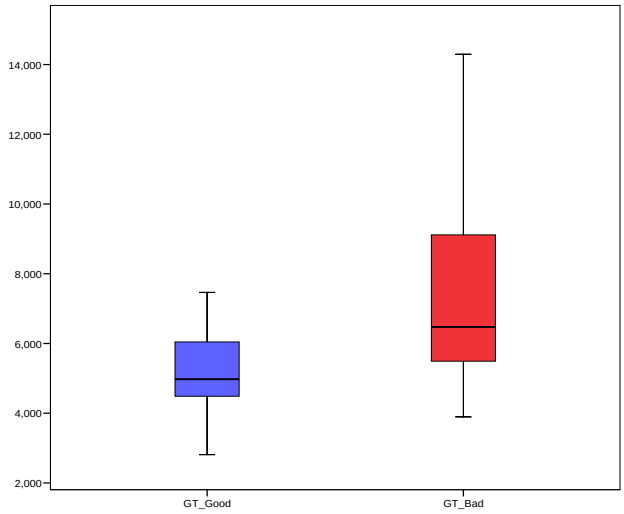

Figure 4.2: Average Gaze Time for Good and Bad Sentences for All Participants (in milliseconds)

The average gaze time for good sentences was 5124.7ms while that for the bad sentences was higher at 7426.6ms. In other words, participants spent, on average, 45% more time looking at bad sentences than good sentences. Spearman's *rho* suggests a medium strength negative correlation between gaze time and sentence quality (r = -.46, $p < 0.01$).

Obviously, some sentences are longer than others. It therefore makes sense to examine the data according to the number of characters per sentence. Looking at gaze time per character, a similar trend is evident in that the bad sentences still had longer gaze time per character than the good sentences (Figure 4.3). Additionally, when the average gaze time per character of all sentences is taken into account (65.89ms), the majority of sentences above this value were rated as bad (65% or 15 of 23).

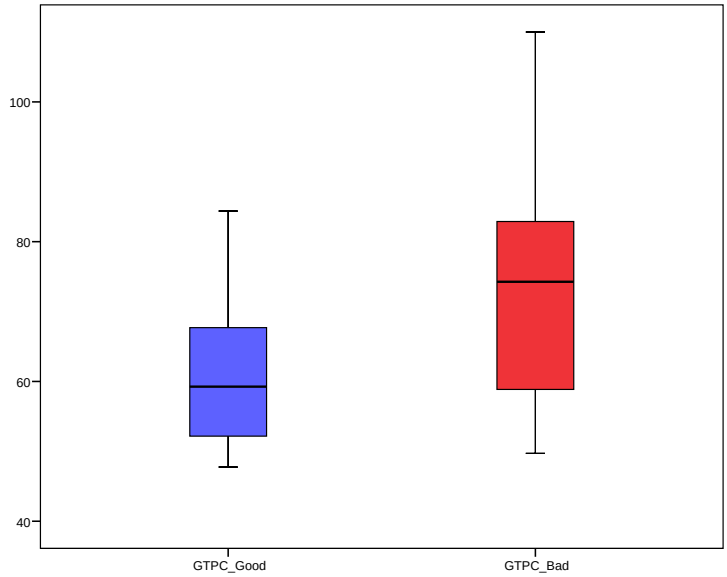

Figure 4.3: Average Gaze Time for Good and Bad Sentences per Character (in milliseconds)

It is interesting to note that the average sentence length for good sentences was 85 characters (median = 78, SD = 28) and bad sentences had a value of 103 characters (median = 97, SD = 40). It was therefore necessary to examine good and bad sentences of similar lengths. By taking the mean character length for all sentences (good and bad sentences combined resulting in a value of 94) and the standard deviation (36), a group of sentences that fall within the standard deviation of the mean can be said to be comparable. Taking ten good and ten bad sentences from this group, the latter still have a higher median gaze time of 7256.9ms to 5190.3ms for good sentences, and a slightly higher fixation count of 89.3 to. 88.1.

The fixation count shows the total number of fixations on a given sentence. Figure 4.4 shows the average fixation count per sentence; a similar trend to that observed in the above figure for average gaze time per sentence is evident, i.e. bad sentences had, on average, more fixations than good sentences. Spearman's *rho* showed a medium strength negative correlation between fixation count and sentence quality ($r = -.47$, $p < 0.01$).

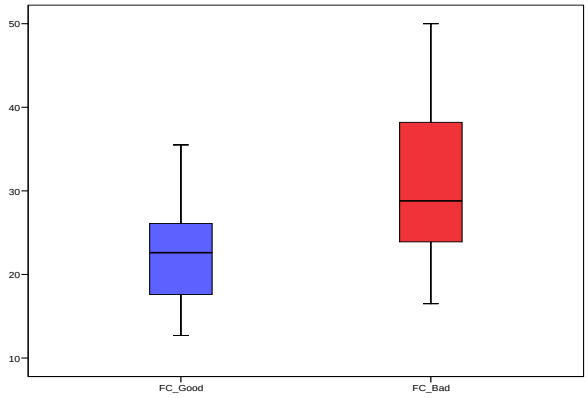

Figure 4.4: Average Fixation Count per Sentence

On examining the median (25.5) of the average fixation count per sentence scores we see that, out of the sentences above the median, 8 sentences were 'good', while 17 were 'bad'.

With regard to fixation count per character, once again there is a negative correlation between this metric and MT quality, as observed above. Additionally, the majority of the sentences that had higher-than-average values were rated as bad (68% or 17 of 25). These results are shown in Figure 4.5:

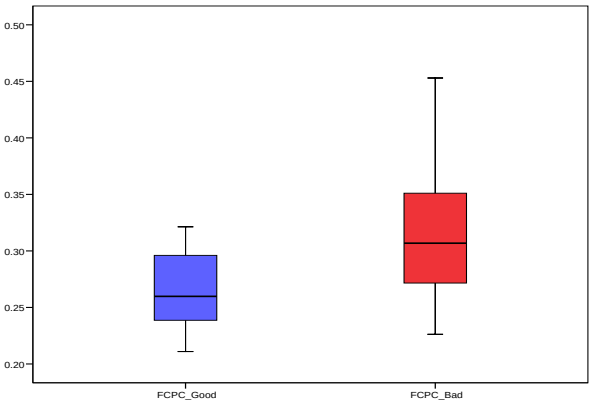

Figure 4.5: Average Fixation Count for Good and Bad Sentences per Character (in milliseconds)

However, it is evident that average fixation durations for good and bad sentences are quite similar, as Figure 4.6 illustrates:

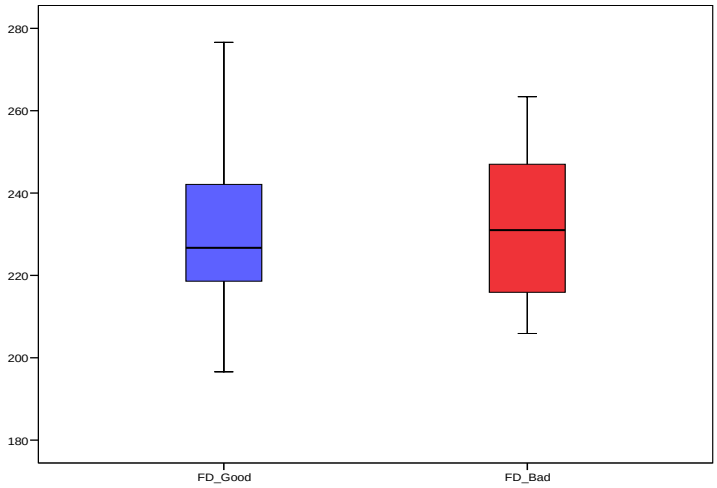

Figure 4.6: Average Fixation Duration (milliseconds) for Good/Bad Sentences for All Participants

The presence of several good sentences among the bad sentences in the highest range of values for average fixation duration is surprising. As already indicated in Chapter Two, an "acclimatisation effect" has been noted before in eye tracking studies (O'Brien 2006), where the initial cognitive effort is higher than for the rest of the task. In light of this, the first five sentences were omitted from analysis.

This elimination had some effect on differentiating the good and bad sentences, though the difference overall was not significant. When fixation duration is viewed per character, the trend is for bad sentences to have longer fixation durations than good ones, but again the differences were found to be non-significant; Figure 4.7 illustrates the effect:

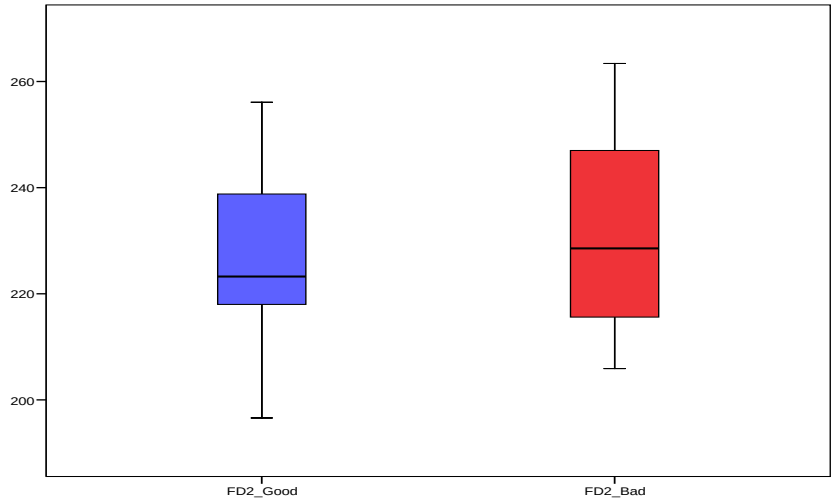

Figure 4.7: Average Fixation Duration (ms) for All Participants from Sentence 6 to 50

.

Lastly, a further measure used to establish a relationship between textual difficulties and cognitive effort is average pupil dilation. On examining the initial results for all sentences across all participants, little difference in average dilation between bad and good sentences was observed (median = 3.83mm and 3.82mm respectively), and no significant difference was found – see Figure 4.8.

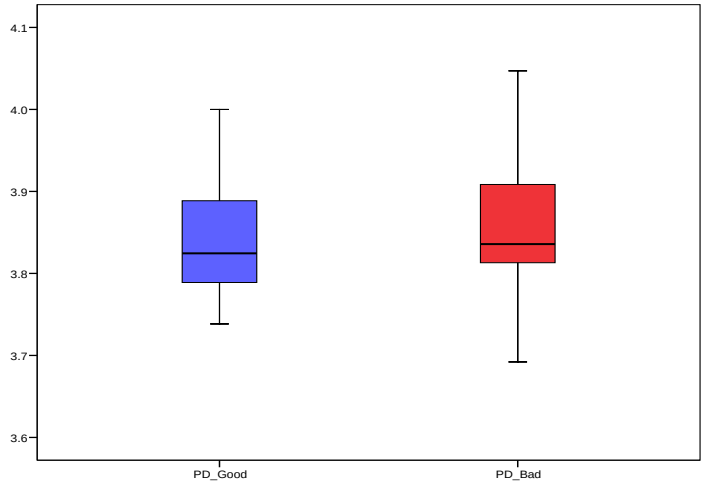

Figure 4.8: Average Pupil Dilation for Good and Bad Sentences (mm)

Given the difficulty in establishing a clear trend in pupil dilation across all participants, it was examined on an intra-subject level motivated by the fact that pupil dilation can vary considerably from person to person. Table 4.1 illustrates that four of the participants (4, 5, 7, and 8) had slightly higher dilation values for bad sentences than good (highlighted in bold) while seven of them either had the same average dilation or had a higher dilation value for good sentences when compared with bad sentences.

Participant	Good Sentence	Bad Sentence
1	3.61	3.61
2	3.91	3.90
3	3.70	3.66
4	**3.32**	**3.37**
5	**2.93**	**2.95**
6	4.02	4.02
7	**3.58**	**3.61**
8	**4.80**	**4.82**
9	3.75	3.70
10	4.87	4.86
11	3.61	3.61

Table 4.1: Average Pupil Dilation (in mm) for Each Participant for Good and Bad Sentences

As previously mentioned, a retrospective protocol with each participant followed the completion of the main task. In this interview, participants were instructed to vocalise their thoughts on their reading patterns as presented to them through the Gaze Replay feature and the retrospective protocol was recorded via Camtasia. As already indicated, the participants' comments were classified as follows: 'All Positive', 'All Negative', 'Mixed', 'Silence', or 'N/A'. Mixed refers to a comment that had both good and bad reports and N/A was assigned when the participant made comments unrelated to the task. In relation to good sentences, 47.2% were met with wholly positive comments. Factoring in the positive remarks in the "Mixed" comments then the participants agreed with the 'good' evaluation in 62.3% of cases. It should be remembered here that at no point were participants aware of the original rating of the sentences. On examining the bad sentences, there was agreement with the initial evaluation in 54.5% of cases. In other words, there was a significant positive correlation where $r = .52$ ($p < 0.01$). Taking into account the mixed comments, as before, a value of 79.2% is reached where $r = .71$ ($p < 0.01$). Good and bad sentences were

met with silence, i.e. no comment of any kind, in 15.3% and 9.6% of cases respectively.

4.5 Lessons Learned

Returning to the initial hypotheses of the pilot study, it was found that:

- Average gaze time was significantly longer for sentences rated as bad;
- Average fixation count was significantly higher for sentences rated as bad;
- There was no significant difference between good and bad sentences for average pupil duration;
- There was no significant difference between pupil dilation for good and bad sentences;
- Participants agreed significantly with the original human evaluation.

It would appear that both gaze time and fixation count were successful indicators of MT output quality whereas the suitability of average fixation duration as a measurement for distinguishing between good and bad MT output requires further investigation. This lack of differentiation in fixation duration has been reflected in other studies in similar contexts. O'Brien (2010) found no significant difference in fixation duration for texts that had been edited using controlled language rules and versions that were uncontrolled. Jakobsen and Jensen (2008) also found insignificant differences in fixation duration across groups in translation process research. Additionally, Van Gog *et al*. (2009, p. 328) suggest that fixation duration may not be an adequate reflection of cognitive load in such a scenario as it represents a different aspect of cognitive processing.

While there have been reports of confounding results using pupil dilation (see Chapter Two), other sources have repeatedly demonstrated an effect on pupil dilation of increased cognitive load (Rayner 1998), and it is evident that further study of pupil dilation as a machine translation evaluation metric is required before coming to any concrete conclusions.

Taking the findings of the pilot study and these issues into account, several changes were made to the research design for the main study. These changes are described in the next chapter.

4.6 Chapter Summary

This chapter described the pilot study, which aimed to validate the proposed methodology and find potential improvements prior to the main study. First of all, the research questions and hypotheses of the pilot study were outlined, highlighting how they differed from those of the main study. The methodology of the pilot study was then described, followed by its results and further discussion. Finally, overall conclusions with recommendations for the next steps of the project were presented, which led to the refinements of the methodology outlined in the next chapter.

Chapter Five

Revised Methods

5.1 Chapter Overview

This chapter recaps on the lessons learned from the pilot study and describes how the issues identified were dealt with in the framework for the main study. First of all, the revisions are detailed. This is followed by a description of the revised experiment design for the main study.

5.2 Revisions

Pupil Dilation

As pointed out in the pilot study, pupil dilation was not found to be an indicator of sentence quality or to correlate well with any other eye-tracking measures. Although this may not be surprising given the reports of confounding results using pupil dilation, other studies have shown a reliable link between pupil dilation and cognitive load (Rayner 1998). Therefore, it is evident that further studies of the use and validity of pupil dilation in this context are needed. It has also been noted that a temporal lag exists between the use of cognitive resources (in the absence or presence of a stimulus) and the resulting pupillary response (Just and Carpenter 1995), which tends to be approximately 1200 milliseconds (Beatty 1982). In an attempt to account for this 'latency' effect in the main study, data were exported from Tobii Studio (via export to .csv format) and the timestamp and matching pupil dilation data were merged into a separate Excel spread sheet, where a latency up to 1500 milliseconds could be accounted for prior to data analysis. In addition to this, as the pilot study found that average values for pupil dilation for sentences did not adequately differentiate between sentences of differing quality, the measure of pupil dilation was changed to percentage change in pupil dilation using the median baseline as described in other eye tracking studies (O'Brien 2006).

Presentation of Stimuli

A 'spill-over' effect was observed in the pilot study whereby participants' pupil dilation on the current sentence was influenced by their reaction to the previous sentence. Similar effects have been observed by Frenck-Mestre (2005). In order to avoid such spill-overs, a blank white slide was shown between each paragraph in the main study; this was not used in the pilot study. These slides were white in colour to minimise the disruption to reading black text on a white background on the following slide.

The white slide appeared automatically after participants pressed the spacebar to indicate that they had finished reading the paragraph in question (the spacebar was also used in this way in the pilot study). The white slide, which served as a barrier between paragraphs was displayed for five seconds. Participants were made aware of this slide in advance and the keystroke-logging data recorded during the sessions using Tobii Studio showed that it did not appear to interfere with the process as no additional key presses, e.g. to move on from the white screen, were recorded.

Lastly, it was also necessary to allow a white space of 5mm between the text and the borders of the screen to capture the data relevant to the texts. The pilot study used individual out-of-context sentences, so space on-screen was not an issue. However, given the length of the paragraphs used in the main study, the distribution of white space became important. Participants read six coherent paragraphs of 150 to 200 words as described in Chapter Two and Three, the indices used in this study require more than 100 words to give an accurate calculation of readability. Due to space constraints and the need to display the text clearly, the paragraphs were split into six slides, separated from each other by the timed blank white slide described above.

Acclimatisation

As discussed previously, the findings of the pilot study indicate the presence of an acclimatisation effect whereby eye tracking measures showed an initial spike while participants were accommodating to the experiment/task required of them and their surroundings (O'Brien 2006, Doherty and O'Brien 2009). As there was a break between the warm-up task and the main task in the pilot study, during which participant were free to ask questions, it was possible that acclimatisation effects were present at the beginning of the main task. Analysis of the pilot data suggests this could be the case.

To address this issue, there was no break between the warm-up and main tasks in the main study in that the first paragraph viewed by the participants was excluded from the analysis of eye-tracking data.

Retrospective Tasks

Finally, it was evident from the analysis of the pilot study data that participants had some difficulties verbalising their thought processes during the retrospective protocols. A possible solution for this would be to prepare a sample protocol in which the researcher comments on fixations, etc., to illustrate the kind of verbalisations to participants might make. Alternatively participants could have been provided with written guidelines.

Both of these approaches might have been suggestive and/or directive, and it was decided to abandon the retrospective protocols in the main study. Instead, participants in the main study were asked to rate sentences from the paragraphs they had seen on-screen for their readability and comprehensibility after the task. This may introduce a bias as participants have seen the sentences previously; however, it was not possible to carry out the evaluation during the task as: (a) the objective was for participants to read the content in a natural way - not to evaluate it at this point, and (b) this was also not possible with the eye-tracking software used in the study. This rating was done using hard-copy print outs of the sentences in question (see Appendix G) and participants also completed a recall test, as defined in Chapter Three (see Appendix F).

5.3 Revised Experiment Design

The main study consisted of three stages (with approximate duration in parentheses):

- Pre-task (5 minutes)
- Main task (10 minutes)
- Post-task (15 minutes)

Pre-Task

Participants were first assigned randomly to the controlled-language or uncontrolled language condition. Upon arrival at the venue where the research was conducted, each participant was seated at a desk adjacent to the eye tracker. Firstly, the experiment and its objectives were described to the participant to ensure it was clear what participation in the experiment would entail. The participant was given time to thoroughly read the instructions and to ask any questions – as required by the university's ethics policy. Once the participant agreed to continue, the Informed Consent Form was signed by the researcher and the participant; both parties retained a copy of this form. A copy of this form can be found in Appendix A. A hard-copy questionnaire (Appendix B) was given to the participant to gather information about the participant, i.e. their name, age, gender, education, professional experience, and knowledge of the domain of anti-virus/technical support documentation. These data were later added to the participant's recording in Tobii Studio by the researcher. These steps also allowed participants time to become as comfortable as possible in the experimental surroundings.

Main Task

Once seated in front of the eye tracker, a calibration lasting 15 seconds involved the participant following a series of red dots on screen to allow the eye tracker to calibrate to the participant's eyes, as ocular features differ from

person to person, and not compensating for differences would result in inaccurate or missing data. Lighting and heating conditions were kept constant so as not to add further environmental variables or cause discomfort to participants. Participants were also instructed not to take caffeine or other strong stimulants up to four hours prior to the experiment as this has been shown to influence eye-tracking data (Michael *et al*. 2008); confirmation that this instruction had been followed was given prior to task commencement.

As described, the paragraphs used in the main study each contained 150 to 200 words. Once again, the guidelines of Gerganov (2007) were adhered to: a screen resolution of 1280 by 1024 pixels, a font style of Tahoma, size 20, double line spacing, and a maximum of 90 characters per line. As already indicated, in order to capture all data belonging to each paragraph, it was ensured that at least 5mm of empty space was left between the text and the borders of the screen.

Post-Task

After completing the main task, the participant moved away from the computer and was invited to sit at another desk. The retrospective task began at this time and consisted of a hard-copy evaluation of the paragraphs seen previously by the participant in the main task. A standard five-point Likert scale[7] was used to measure the participant's opinion of the readability and comprehensibility of the text; the format of the scale was:

1 = strongly disagree;
2 = disagree;
3 = neither agree nor disagree;
4 = agree;
5 = strongly agree.

Participants were asked to evaluate each sentence individually, in the

[7] One of the most common ways of measuring "a person's feelings or attitudes toward another person, event, or phenomenon" (Frey *et al*. 1999, p. 103) on an interval scale.

order in which they appeared in the paragraphs originally read on screen. Participants remained in their previously assigned conditions (either uncontrolled or controlled) and so saw the same sentences as they had previously seen on screen. Participants were asked to indicate their level of agreement (from strongly agree to strongly disagree) with the statement that a given sentence was readable and comprehensible. Definitions of both concepts were given both on the cover page of the evaluation booklet, and at the bottom of each subsequent page to serve as a reminder. Figure 5.1 provides a sample; the full versions can be found in Appendix G.

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)

Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

In other words, the higher the number on the scale, the better you could ***read*** and ***comprehend*** the sentence.

Simply mark the number relating to the sentence to judge how ***readable*** and ***comprehensible*** the sentence is. Here is an example:

1. Avant de passer à l'étape suivante, assurez-vous que le logiciel est mis à jour.

This sentence is readable.
1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

1 2 3 4 5

This sentence is comprehensible.

1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

1 2 3 4 5

Don't worry! The legend and explanations will appear on each page, so you won't need to memorise them. Take as much time as you need.

Figure 5.1: Instructions from Evaluation

Finally, a hard copy of the recall test described in the previous chapter was administered to test for comprehension. An example can be seen below in Figure 5.2, and the full version can be found in Appendix F. Upon completion of the text, participants were given the opportunity to provide any feedback or ask any questions, then thanked for their participation in the study and escorted out of the research lab.

	1. À propos de Symantec AntiVirus		
1.1	After reading this paragraph, do you understand the options for installing Symantec AntiVirus?	Yes	No
1.2	Votre administrateur exécute des analyses sur votre ordinateur et peut configurer des _____ _____.'		
1.3	What are the consequences of an administrator managing the installation?		

Figure 5.2: Sample from Recall Test

5.4 Data Preparation

5.4.1 Data Formats

Data from the readability indices, recall test, and the human evaluation output (readability and comprehensibility) were collated and saved in Microsoft Excel's .xls format. Eye tracking data were exported from Tobii Studio to Excel, and were then edited to isolate the required data. Lastly, all data were then manually imported into SPSS (version 16) for statistical analysis.

5.4.2 Units of Measurement

Table 5.1 lists units of measurement for each variable investigated in this study. The scores for the human evaluation measures of readability and comprehensibility (rows 11 and 12) were converted to a standard score (z-score) as they rely on a constructed test where the possible maximum value differs for each condition due to the number of sentences being higher in the uncontrolled condition than the controlled. The z-score is the number of standard deviations a particular score lies above or below the sample mean. Conversion to z-scores allows data to be compared as standard deviations are a universal scale of measurement.

	Variable	Measurement	Form
1	Flesch	Interval	Numerical Raw Score
2	LIX	Interval	Numerical Raw Score
3	GTM	Interval	Numerical Raw Score
4	BLEU	Interval	Numerical Raw Score
5	TER	Interval	Numerical Raw Score
6	Observation Length	Interval	Millisecond
7	Fixation Count	Interval	Numerical Raw Score
8	Fixation Length	Interval	Millisecond
9	Percentage Change in Pupil Dilation	Interval	Percentage
10	Regressions	Interval	Numerical Raw Score
11	Readability Evaluation	Interval	Z-Score
12	Comprehensibility Evaluation	Interval	Z-Score
13	Recall	Interval	Numerical Raw Score

Table 5.1: Measurement and Form of Each Variable

5.5 Data Quality

It has been well documented that eye tracking data are sensitive to many participant and environmental factors (O'Brien 2009). Therefore, it is paramount to ascertain that the quality of the data captured by the eye tracker is adequate for data analysis and interpretation. As described in Chapter Three, several precautions were taken to minimise poor data capture. However, it is still advisable to ensure bad quality data are isolated. Mean fixation length, gaze time on screen (GTS), sample rate, and total task time were used as criteria on which the data could be filtered. Participants whose data did not meet two or more of the criteria described below were not used in further data analysis. It should be noted that for later data analysis, participants were renumbered after this assessment of quality. Therefore, participants who did not meet the established criterion appear towards the bottom of the table in each case. Table 5.2 provides data for each criterion, highlighting violations in grey.

Participant	Total Task Time (sec)	Mean Fixation Length (ms)	Sample Rate	GTS (%)
1	541.753	373.333	98	88.90
2	399.579	311.666	93	87.73
3	326.579	201.666	67	61.89
4	269.365	210.998	77	75.44
5	524.365	293.333	83	78.63
6	502.236	321.666	79	89.84
7	498.265	231.666	91	71.57
8	385.65	246.833	78	66.18
9	423.12	267.333	59	87.33
10	475.586	249.333	89	62.17
11	284.514	253.333	98	79.54
12	515.561	263.333	94	85.33
13	520.909	243.333	92	76.54
14	514.322	258.833	97	81.90
15	489.258	241.666	98	78.76
16	448.36	239.333	93	87.45
17	305.254	210.166	59	62.01
18	379.113	319.198	83	80.54
19	301.221	275.166	80	71.58
20	307.594	201.666	67	63.63
~~**21**~~	201.358	231.981	31	43.32
~~**22**~~	365.21	198.233	14	72.91
~~**23**~~	459.255	177.233	44	22.04
~~**24**~~	501.269	189.233	14	70.95
~~**25**~~	191.568	112.333	23	12.31

Table 5.2: Estimation of Quality of Eye Tracking Data

Total Task Time

Total task time as measured in Tobii Studio represents the time from when the researcher began recording until the participant completed reading the final text. In the current study, the mean for task time was calculated (mean = 413.737 seconds) and participants whose score was more than one standard deviation from the mean (SD = 107.279 seconds) were flagged for potentially problematic data, as these participants may have not completed the task and may have merely sought to end the experiment early, or they may have spent more time than needed to complete the task e.g. because they may have felt they were going too fast (GTS as described below addresses this further). Six participants did not meet the criterion of falling within one standard deviation of the mean score for total take time, and were flagged.

Mean Fixation Duration

Average fixation duration has been noted to range between 225 and 400ms for reading (Rayner 1998) and has been used as a means of ascertaining data quality in other eye tracking studies (Pavlovic and Jensen 2009, Hvelplund 2011). In the current study, the mean fixation duration was calculated for each participant by dividing the sum of their fixation duration over the entire task by their fixation count. In other studies, a number of participants' recordings were excluded due to very short mean fixation durations of less than 200ms (O'Brien 2009, Pavlovic and Jensen 2009). In this study, the mean fixation duration for all participants was 244.91 milliseconds. Seven participants fell below the minimum threshold of 225 milliseconds and were flagged. For participant 25 this second violation resulted in exclusion.

Sample Rate

Tobii Studio uses its own metric for measuring the quality of a recording: sample rate. Sample rate is calculated using the number of samples the eye tracker could identify during recording. A value of 100% equates to both eyes

being found throughout the recording. 50% means that only one eye could be found or both eyes for half of the recording time. Time spent looking away from the screen results in a decrease of these values. Tobii Studio allows for one or both eyes to be used in its sample rating, in this case, both eyes were selected and a threshold of 50% was used. The reasoning behind this value is that, as stated, a value lower than 50% indicates that only one eye could be identified during the full recording, or possibly that both eyes were identified for half of the recording time. This would indicate possible inaccuracies in the recording, and so such instances should be flagged. Five participants did not meet this threshold and this represented exclusion for participants 21, 22, 23 and 24, and further justification for the removal of participant 25.

Gaze Time on Screen (GTS)

In addition to the above, gaze time on screen was calculated independently of Tobii Studio to ensure a consistent quality assessment of the recordings. GTS was calculated for each participant by dividing total observation length by total number of fixations and multiplying this value by 100 to attain a percentage value (see Hvelplund 2011). Once again, the mean for all participants was found (mean = 72.76%) and values falling outside one standard deviation (SD = 19.44%) were flagged. This provided further evidence that data from participants 21, 23, and 25 should be excluded.

In summary, by combining known estimators of data quality with the existing measure used by Tobii Studio, the risk of poor quality data being used in further analysis and interpretation could be reduced. As with other studies in eye tracking, and indeed in the pilot study, participants' recordings were removed due to poor quality and other problems. In the current study, a total of five were excluded from data analysis out of a possible twenty-five. In addition, for measurement of pupil dilation, a latency effect of 1500ms, which has been documented in the literature (Hyönä *et al*. 1995), was accounted for. As Tobii Studio enters -1 values into the data output when it loses contact with the pupil, these values were removed and replaced with a blank cell so as not to interfere with the calculation of means, medians, or percentage change in pupil dilation.

5.6 Methods of Analysis

Table 5.3 provides an overview of all variables investigated in this study. They have been grouped according to their conceptual nature and sequence of their collection:

Group A: Textual	Group B: Eye Tracking	Group C: Evaluation
Flesch	Observation Length	Readability Evaluation
LIX	Fixation Count	Comprehensibility Evaluation
GTM	Fixation Length	Recall
BLEU	Percentage Change in Pupil Dilation	
TER	Regressions	
	Regression Distance	

Table 5.3: Overview of Variable Groupings

It should be highlighted that the uncontrolled and controlled groups (of participants) represent two conditions in the experiment design. In the sense of each group being a condition, the term uncontrolled and controlled condition are used. These conditions/groups refer to the controlled language aspect of the study, i.e. the uncontrolled-language condition and the controlled-language condition, denoting the nature of the paragraphs shown to each group of participants. Therefore, the term controlled condition/group should not be confused with an experimental design in which a 'control' or placebo group is used, as this is not the case in the current study. In correlational analysis there are no independent variables or dependent variables, rather two (or more) variables that have a relationship with one another in some way, and causality is not taken into account. In analysis of variance, the test variable is the score, e.g. Flesch, which is examined for a difference between the grouping variable, i.e. the two conditions above. In this sense, the three groupings described above do not contain independent or dependent variables per se. Rather, the conditions (uncontrolled and controlled) could be said to be the dependent variable, as is the case in the multiple regression analysis described below. The following subsections describe each of the statistical analyses that were carried out on the data.

5.6.1 Correlation Coefficients

"A correlation coefficient is a numerical index that indicates the strength and direction of a relationship between variables" (Howitt and Cramer 2008, p. 76). The most commonly used coefficient is Pearson's *r*, while other variants such as Spearman's *rho* are more appropriate under certain circumstances, e.g. when using non-parametric data. The measure of the correlation ranges from -1.00 to +1.00 and the greater the value, the greater the strength of the correlation. Positive values indicate a positive correlation, i.e. as one variable increases, so does its correlate; negative values indicate a negative relationship as one variable increases, its correlate decreases. A value of +/-1.00 indicates a perfect association of two variables where an exact linear relationship exists, while a value of 0.00 points to values having a random relationship, where no straight line could be drawn, for example, by using a scatter plot.

5.6.2 Independent Samples T-Test

The independent samples t-test is also known as an unrelated or uncorrelated t-test, and is used to determine if the means of two scores of the same variable differ between two groups/conditions. When a significant difference is found this indicates that the difference between the two groups/conditions is neither due to chance nor sampling (see below).

5.6.3 Analysis of Variance (ANOVA)

ANOVA is an extension of the t-test in that it allows for more than two groups/conditions of participants. It compares the variation in the means between each condition with the variation within each condition and gives an overall value of variance called the *F*-ratio.

5.6.4 Multivariate Analysis of Variance (MANOVA)

MANOVA is a form of ANOVA that is applied when analysing several groups/conditions at once. The analysis examines if the means of the independent variables differ significantly on the combined dependent variables. Therefore, it avoids overestimation of significance due to carrying out multiple significance tests on the same data, i.e. multiple ANOVAs. If the MANOVA reveals a significant result, then it is necessary to test the significance of the individual variables by means of individual ANOVAs or t-tests. As this study contains two conditions, the independent samples t-test is most appropriate given that it is the convention for the comparison of just two groups/conditions, whereas ANOVAs are more suited to more groups/conditions, yet both methods would find the same result (Howitt and Cramer 2008, p. 165).

5.6.5 Multiple Regression Analysis

Multiple regression analysis identifies predictors amongst the independent variables in relation to a particular dependent variable by means of statistical criteria. The preferred way of identifying predictors on the basis of empirical statistical data is stepwise regression. Other options such as hierarchical regression were not appropriate in the current study as the literature does not indicate consistent and significant predictors of readability and/or comprehensibility. Stepwise regression takes one variable at a time and identifies how much of the variance in the dependent variable is accounted for by this variable. This process continues until no significant increase is observed in the models, therefore excluding variables that did not account for a significant amount of variance. The first predictor, which has been identified as the strongest, accounts for most variance in the dependent variable. However, further predictors are also worthy of note and help to further explain interaction between variables.

The strength of this method is that it is ideal for identifying the smallest number of predictors of a particular variable of interest, e.g. Flesch scores. In the case of this study, several variables have been included as possible predictors of

readability and, due to the experimental design, such inclusions are useful in further explanation of the findings.

When interpreting results of multiple regression it is important to be mindful of variables that have a high correlation with each other, i.e. show multicollinearity/covariance. When used together it can be the case that these predictors can skew the model. Therefore, it is necessary to ensure that identifiable predictor variables have no such relationship and this can be achieved by correlational analysis, and examination of interactions provided in the regression model.

5.6.6 Significance Testing & Outliers

As described earlier, the cut off for statistical significance is 0.05, i.e. there is a 5% chance that a result is due to chance. The hypotheses of this study do not stipulate the direction of the relationships between the measured variables, that is, the hypotheses are non-directional and require a two-tailed test of significance to be used. Asterisks are used to indicate levels of significance, where * equates to .05 and ** to .01; these figures are usually given in parentheses so as to be consistent with the literature.

Lastly, for the sake of clarity outliers are observations where the score is numerically distant from the sample mean. In the current study they were identified as being more than two standard deviations above or below the sample mean, and are identified by a star on graphical representations with the number corresponding to the participant/observation number.

5.7 Procedure

Several types of analysis were conducted on each of the groupings:

1. MANOVA to test for significant differences between the controlled and uncontrolled condition in terms of each grouping of dependent variables;
2. Upon finding a significant difference, an independent samples t-test on individual variables;
3. A correlational analysis to examine the relationship between variables to supplement and further explain the above results.

Lastly, a multiple regression analysis was employed to derive more practical findings from the data, on the basis that if predictors of readability could be identified then they could be used to inform recommendations for improving readability and comprehensibility in writing practices and the use of CL in an MT context in general.

Part III:

Data Analysis

Chapter Six:

Results and Discussion

6.1 Chapter Overview

This chapter presents results for each grouping of variables identified and described in the previous chapter. The results for each analysis are presented in the order of: a MANOVA test for significant differences between groups/conditions, individual independent samples t-tests, and correlational analyses. There are inevitable problems in presenting results in sequence and, where appropriate, indications will be given when points are discussed in or relate to other sections in the chapter. The structure for each grouping will follow the formula:

I. Introduction to Grouping
II. Overview of Data in Tabular Format
III. Description of Data
IV. Interpretation of Data with Graphical Elements
V. Internal and External Interactions of Groupings
VI. Discussion (Comparative and Topic-Oriented)
VII. Grouping Summary

The textual variables of the error classification analysis, the Flesch and LIX readability indices, and the automatic evaluation metrics comprise Grouping A. Grouping B contains the eye tracking metrics of observation length, fixation count, fixation length, pupil dilation, and regressions. Grouping C concerns itself with the human evaluation variables obtained from the post-task readability and comprehensibility evaluation, and the post-task recall test. In addition, the multiple regression analysis identifies predictors of readability and comprehensibility. Lastly, the results of the study are tested against the hypotheses and inform a chapter summary at the end.

Grouping A: Textual Variables

6.2.1 Section Overview

The first grouping of variables consists of textual variables which were calculated independently of each other. The variables in question are: Flesch and LIX scores and the automatic evaluation metrics (AEMs) of GTM, BLEU, and TER. Significant differences between conditions on these variables would point to the implementation of the CL rule set having an effect, whether positive or negative, on the texts from the point of view of the objective textual measures. High levels of correlation between these variables would indicate consistency between these metrics and possibly indicate construct validity and reliability of the measurements, while differences in their results would highlight discrepancies between metrics and present issues worthy of further investigation.

A MANOVA showed no significant difference between conditions where Pillai's $F = 2.266$, $df = 5.0$, $p < 0.05$, partial $\eta^2 = .654$. However, given the independence of the metrics, it was still advisable to examine each in closer detail, using the statistical techniques referred to in Chapter Five.

A closer qualitative analysis of the uncontrolled and controlled MT output was also in order, to see what, if any, differences arose in the two outputs. This kind of analysis could help to explain results from statistical analyses, and perhaps later human evaluations. It was decided to base this textual analysis on the kind of error analysis often adopted in MT research.

This section first describes the analysis of the errors produced in the output of the MT systems using a typography proposed by Viler *et al*. (2006) and discusses the differences between the controlled and uncontrolled output in this regard. Following this, the variables of readability indices (Flesch and LIX) and AEMs are analysed by means of descriptive and inferential statistics. Followed by further discussion of results and a section summary.

6.2.2 Error Categorisation

Comparison of different MT systems is usually achieved on the basis of automatic metrics that measure, for instance, edit distance between the raw MT output and one or several gold standard references, at various levels ranging from unigrams to overall document level. As previously described, GTM, BLEU, and TER are common examples of such AEMs. While these metrics provide a fast and resource-light means of evaluation, the interpretation of the measures and their scores does not always correlate with human evaluation or other AEMs (see Section 7.1.3). Human analysis of the output of the MT system is therefore useful in order to build a more thorough picture of the quality of the output and isolate errors and other issues, especially during the later development processes and prior to dissemination or post-editing.

Using the classification model proposed by Vilar *et al.* (2006), the errors in the MT output for the controlled and uncontrolled texts were identified and categorised – see Figure 6.1 and Table 6.1. This model was used due to its widespread application in MT evaluation in research and development scenarios, e.g. Parton and McKeown 2010, Propvić and Burchardt 2011, and the evaluation was carried out by the researcher. Such an evaluation is, of course, rather subjective. However, it is hoped that combined with the other findings in this chapter, the results in this section provide complementary insights into the quality of the MT output.

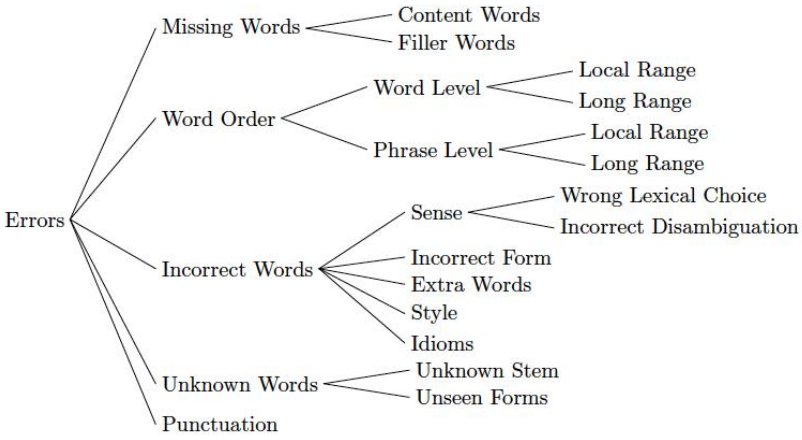

Figure 6.1: Classification of Translation Errors after Vilar *et al.* (2006)

The classification typology adopts a hierarchical structure – see Figure 6.1. The first tier contains the five main categories of errors: missing words, word order, incorrect words, unknown words, and punctuation. Table 6.1 provides descriptions for each category and its subcategories that were used in the error classification.

Type	Sub-type	Description
1. Missing Word		**Word missing in generated sentence**
	Content Word	Missing word necessary for meaning
	Filler Word	Missing word necessary for grammatical correctness
2. Word Order		**Word placed incorrectly in generated sentence**
	Word Level	Necessary to move individual words which are independent of each other
	Phrase Level	Necessary to move blocks of consecutive words together
3. Incorrect Word		**When the system is unable to find the correct translation**
	Sense	The incorrect word disrupts meaning when incorrect word is chosen, or when the system is not able to understand the source word in the given context
	Incorrect Form	Correct translation found but the system is unable to produce its correct form
	Extra Word	Additional words added to the translation which are not present in the source
	Style	Bad choice of words for the translation given its context
	Idioms	Idioms the system does not know and translates as normal text
4. Unknown Word		**Words the system does not know**
	Unknown Stem	Truly unknown words
	Unseen Forms	Stem known but form unfamiliar
5. Punctuation		Errors in punctuation

Table 6.1: Descriptions of Main Error Categories

In many cases in the current study, several errors were identified in a single sentence and rectifying one error, e.g. local word order, could also result in resolving other identified errors in the sentence. In such cases, all errors were taken into account. Therefore, the number of errors may be somewhat high for such a short series of texts. Table 6.2 summarises the errors identified in both conditions.

Type	**Sub-type**	**Uncontrolled**	**Controlled**
Word Count		**1450**	**1361**
Missing Word		**93** (17.12%)	**64** (13.33%)
	Content Word	21	17
	Filler Word	72	47
Word Order		**64** (11.78%)	**64** (13.33%)
	Local Word Order	64	64
Incorrect Word		**371** (68.32%)	**339** (70.62%)
	Sense		
	Wrong Lexical Choice	245	254
	Incorrect Form		
	Verb	11	27
	Person	0	0
	Gender	7	10
	Number	26	25
	Extra Word	82	23
Unknown Word		**15**(2.78%)	**13**(2.72%)
Total		**543**	**480**

Table 6.2: Errors for Both Conditions

Overall, Table 6.2 shows that the total number of errors for each of the outputs was somewhat similar: 543 for the uncontrolled (or 37.45 errors per 100 words), and 480 for the controlled (35.27 errors per 100 words). Additionally, the distribution of errors across the categories is also rather similar. Missing words accounted for 17.12% of errors in the uncontrolled output, and 13.33% for the controlled. The latter had a lower number of missing words, both content and filler. For the uncontrolled text, in this category, prepositions and articles were particularly problematic. Errors in word order accounted for 13.33% in the controlled output, greater than that of 11.78% in the uncontrolled, although the two texts had exactly the same number of absolute errors in this category. No errors occurred beyond the local level, which is unsurprising given the sentences in the texts contained only one clause in the majority of cases. In both outputs, incorrect words accounted for the vast

majority of errors with a similar number (and percentage) for both conditions: 371 (68.32%) in the uncontrolled, and 339 (70.62%) in the controlled. Of these errors, under the subcategory of *Sense*, wrong lexical choices were the leading cause of errors, again similar for both outputs: 245 for the uncontrolled, 254 for the controlled. Further into this category greater differences can be seen: the uncontrolled text had fewer errors in verbs (11 to 27) and gender (7 to 10), similar errors in number (26 to 25), yet far more errors in adding extra words (82 to 23). Lastly, unknown words accounted for 2.78% and 2.72% of errors in the uncontrolled and controlled texts respectively with a similar number of cases once again. Neither output contained any errors in punctuation. Once again this is unsurprising considering the length of most sentences and the presence of one clause per sentence in most cases. Regardless, this result shows that both MT systems dealt perfectly with punctuation. The overall proportions of errors are represented graphically in Figure 6.2.

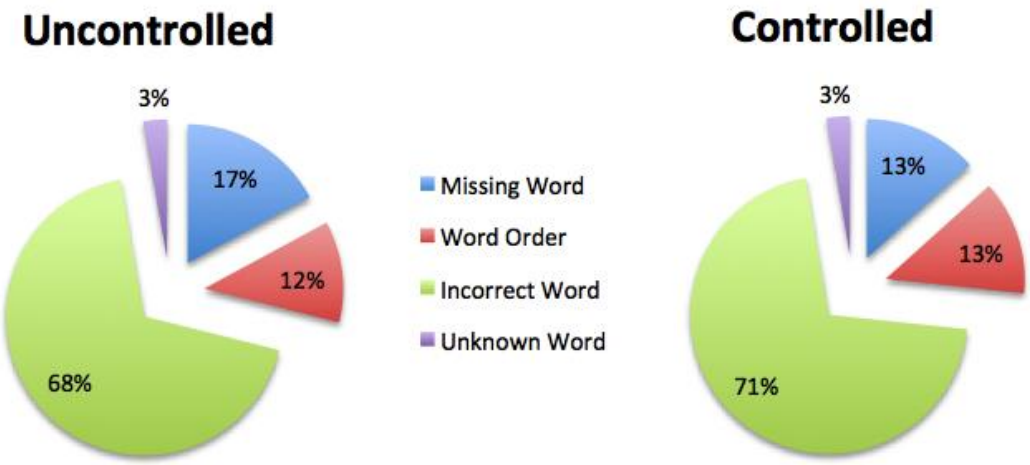

Figure 6.2: Errors for Both Conditions

These findings indicate somewhat of a difference between the conditions in terms of the number of errors present in the output and the distribution across the types and sub-types. It should be noted that the distribution of errors is similar to that of Vilar *et al*. (2006) in that *Incorrect Word* errors are substantially greater than any other categories. Overall, it is interesting to highlight the general trends in that the controlled system had much more

success in retaining words in the target, where the uncontrolled tended to miss words, mostly in prepositional phrases. The controlled system was slightly more problematic with verb conjugation and tended to leave verbs in their infinitive form. The uncontrolled system also had a recurring problem with inserting additional unnecessary words, which was much more pronounced than in the controlled system.

These results do not differ greatly to those of O'Brien (2010) who also used the error analysis categorisation proposed by Vilar *et al.* (2006) and found that readability (and acceptability) increased marginally for the controlled version of the 'difficult' texts she investigated, but not for the text rated by Flesch to be easier. A similar spread, yet smaller number of errors, was also found. Similarly, De Preux (2005) found that the number of errors did not decrease in a study of CL but that the severity of these errors was reduced by the application of the CL.

6.2.3 Readability Indices

An overview of the raw scores provided by the Flesch and LIX readability indices is given in Table 6.3 and shows the scores for each paragraph. As described earlier, texts selected were between 150 to 200 words to overcome issues with calculating accurate scores for sub-100 word texts, and also to provide a more realistic reading experience for the participants. For the Flesch measure a higher score indicates greater readability. To serve as examples, scores of 70 and above indicate that a text should be easily readable for a school-going teenager, while scores of 30 and under indicate texts that are best suited to university graduates. For LIX, higher scores represent reduced readability: scores of 30 and under are deemed to indicate texts that are easy to read and typically represent children's literature; scores of 40 to 50 are typical of a daily newspaper; while scores greater than 60 indicate texts that pose the greatest amount of difficulty. Such scores are usually given to highly specific genres such as legal and technical texts.

	Flesch			**LIX**	
Paragraph	*Uncontrolled*	*Controlled*		*Uncontrolled*	*Controlled*
1	25	31		67	60
2	23	22		73	73
3	52	54		56	55
4	44	46		60	61
5	44	45		62	60
6	44	44		63	62

Table 6.3: Flesch and LIX Scores for Both Conditions

Firstly, there is a difference between conditions for Flesch scores in all but one instance (paragraph 6) where the score is the same. The difference between the conditions for Flesch scores is not statistically significant (t = -.24, df = 10, p = 0.81). There is, however, an increased mean for the controlled condition: 38.67 (SD = 11.79) for the uncontrolled and 40.33 for the controlled (SD = 11.63), where a higher Flesch score equates to better readability. The uncontrolled condition has a range of 29 (min. = 23, max. = 52) and a median of 44 and the controlled condition has a range of 32 (min. = 22, max. = 54) and a

median of 44.5. Figure 6.3 shows this information and also shows the presence of two outliers.

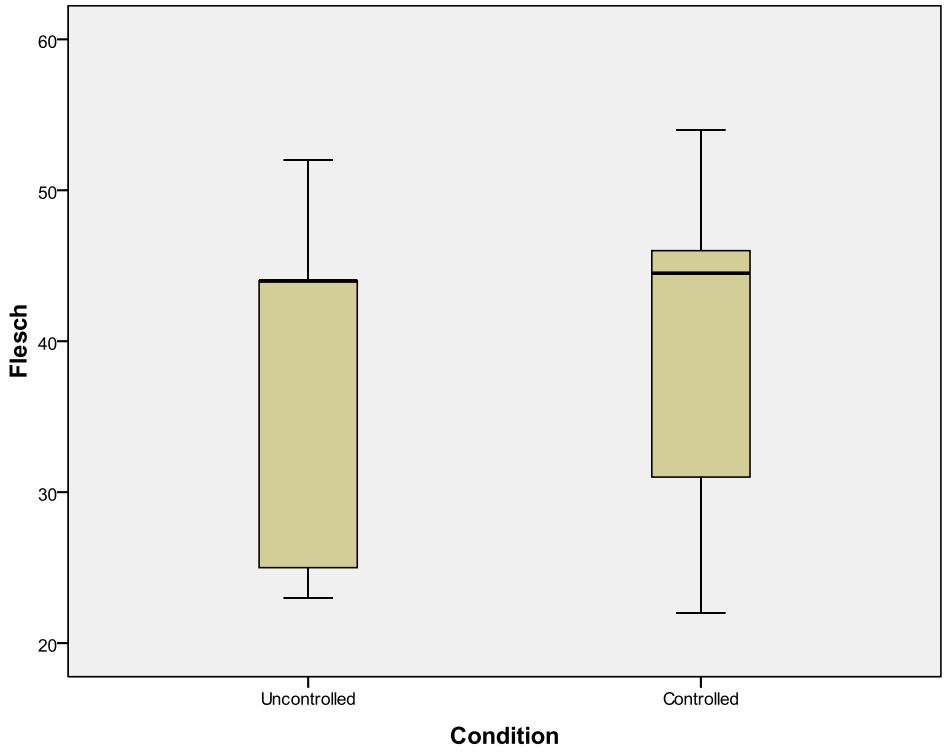

Figure 6.3: Flesch Scores for Both Conditions

Secondly, LIX shows similar results where the scores fall in the controlled condition (lower scores equate to higher readability) except in paragraph 4 where the score actually increases. Once again, the difference between the conditions for LIX scores is not statistically significant (t = .48, df = 10, p = .63). The uncontrolled text had a mean of 63.5 (SD = 5.89), a median of 62.2 with a range of 17 (min. = 56, max. = 73), whereas the controlled text had a mean of 61.83 (SD = 5.98), a median of 60.5, and a range of 18 (min. = 55, max. = 73). Figure 6.4 shows these values and also indicates an outlier. Once removed, the difference remains insignificant where t = -.377, df = 9, p = 0.63, which indicates that there was no significant improvement in LIX scores due to the application of the CL rule set.

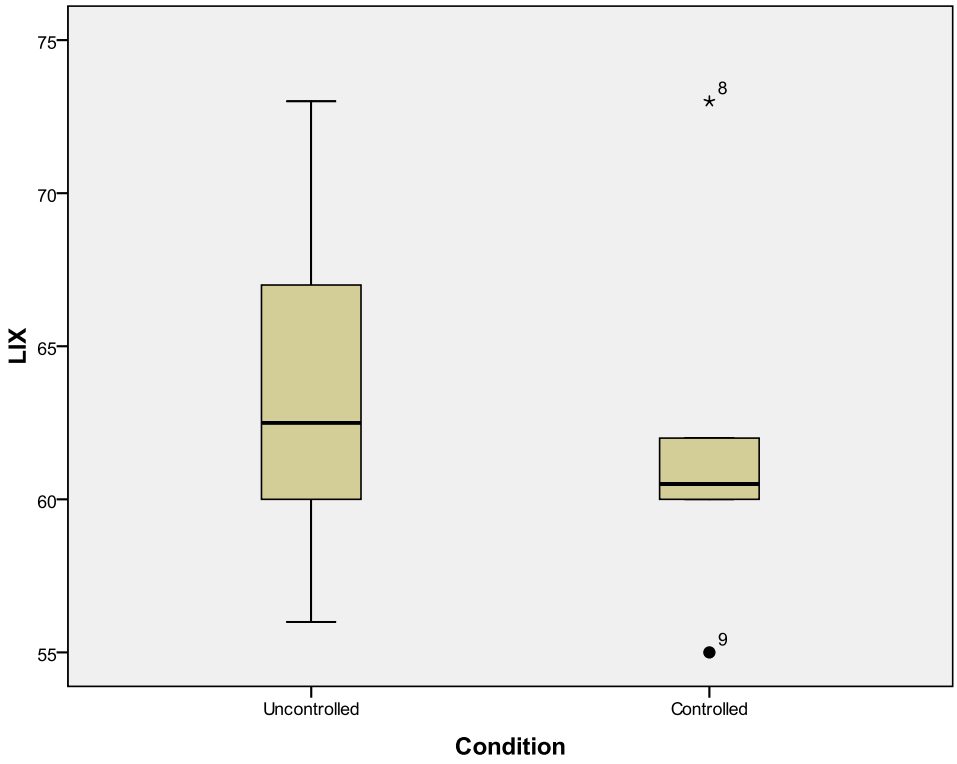

Figure 6.4: LIX Scores for Both Conditions

Lastly, it is also the case that Flesch and LIX scores showed a strong negative correlation r = -.89, p = .001, i.e. as Flesch scores increase, LIX scores decrease. Figure 6.5 shows both scores on a scatter plot and illustrates the correlation.

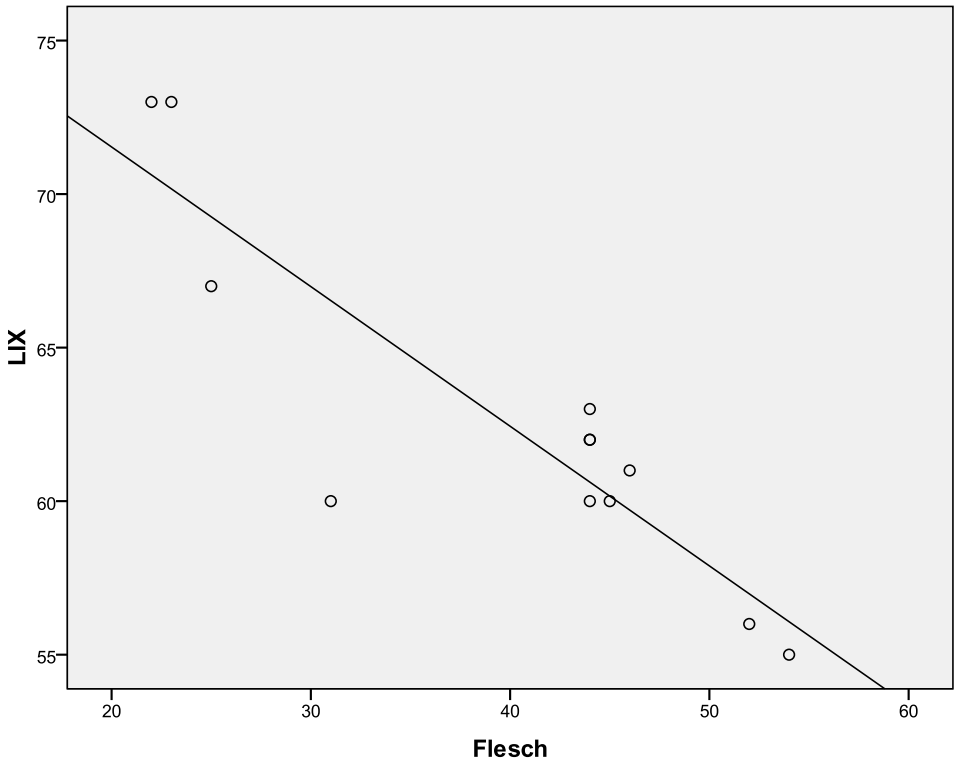

Figure 6.5: Correlation between Flesch and LIX

In similar studies of controlled language and readability, Cadwell (2008) and later O'Brien (2010) found little increase in readability as measured by Flesch scores when a CL rule set was implemented. The findings here largely support these reports. However, O'Brien's study (2010) also found that the improvement of readability was more prevalent in the text rated as very difficult (17.8, which improved to 19.2 when controlled), than the easier text (71.1, which only marginally improved to 71.3), both of which represent a margin improvement in readability. In other words, CL appears to improve readability, as measured by Flesch, for texts rated as 'difficult' more than those already rated as 'easy' to read. This stands to reason as, in theory, easier texts would not benefit as much from the use of a CL rule set in that the text may already be as readable and unambiguous etc. as possible. In the case of the current study, the moderately difficult texts in terms of Flesch/LIX scores may be a reason for the non-significant result, which supports the findings of Cadwell (2008) and O'Brien (2010). Lastly, Jensen (2009) also found a similarly strong correlation between these two measures, which was also the case here.

6.2.4 Automatic Evaluation Metrics (AEMs)

This section concerns itself with the evaluation of the MT output for both conditions using automatic evaluation metrics (AEMs). As described in Chapter Two, despite problems associated with them, AEMs provide a low-resource solution to MT evaluation where a fast and cheap analysis of the MT output can provide indications as to improvements in the development of the MT system. However, AEMs rely on largely superficial means of textual analysis, e.g. edit distance. When used in conjunction with other methods of evaluation such as usability testing, or human evaluation, or when they are based on multiple reference translations of high quality, AEMs can be more accurate and have been shown to correlate well with human judgments. The correlation between AEMs and human evaluation scores features extensively in the literature (Papineni *et al.* 2002, Turian *et al.* 2003, Snover *et al*. 2006, Callison-Burch *et al*. 2008). Three AEMs were used in the current research: GTM, BLEU, and TER, and this section provides the results for each.

Each target sentence was tagged and aligned with the reference sentence and all paired sentences were input into an algorithm, which automatically generated the GTM, BLEU, and TER scores. The following example demonstrates how the sentences were paired for the AEM evaluation:

Example 1:
<segment1> This is a target sentence. </segment1>
<reference1> This is the paired reference sentence. </reference1>

Scores were averaged for each paragraph and for document level (where 'document' is understood here as referring to all six paragraphs together in the same document) as sentence-level evaluation is not the intended use of metrics such as BLEU. This paragraph-level analysis was also chosen for comparative purposes given that other variables (e.g. Flesch and LIX scores) in the study are only accurate at paragraph level. In the following, the results for each metric (overall system/document level scores) are presented and several discussion points are noted concerning the overall results of the AEMs and how they relate

to the other variables presented thus far. It should also be noted that, as with the readability indices, mean scores in this section reflect each paragraph as possible acclimatisation effects are not present for textual metrics, whereas for eye tracking metrics, the results from the first paragraph were not included due to acclimatisation of participants to the task.

6.2.4.1 System Performance

Focusing on the overall system level provides an overview of how each iteration of the MT system, uncontrolled and controlled, was rated by each AEM. Table 6.4 presents the system scores as rated by each metric. As a reminder, GTM and BLEU equate higher scores (from 0 to 1) with higher similarity to the human reference translation. Conversely, increases in TER scores represent the number of edits required to change the MT output to match the reference. Table 6.4 shoes that the uncontrolled system performed better on all three metrics.

	GTM	BLEU	TER
Uncontrolled	.514	.452	.507
Controlled	.451	.366	.576

Table 6.4: System Scores for Each AEM

6.2.4.2 GTM

As previously stated, GTM measures the target translation against the reference in terms of similarity. Table 6.5 shows that the uncontrolled translation was generally more similar to its reference than the controlled. An independent samples t-test found the difference to be insignificant where the uncontrolled condition (mean = 0.534, SD = 0.190) had a slightly better score than the controlled (mean = 0.488, SD = 0.209) as df = 118, t = 1.262, p = .209. Figure 6.6 gives these scores per paragraph.

Paragraph	1	2	3	4	5	6	Mean
Uncontrolled	0.584	0.382	0.476	0.57	0.592	0.629	0.534
Controlled	0.590	0.458	0.461	0.526	0.341	0.526	0.488

Table 6.5: GTM Scores per Paragraph with Mean

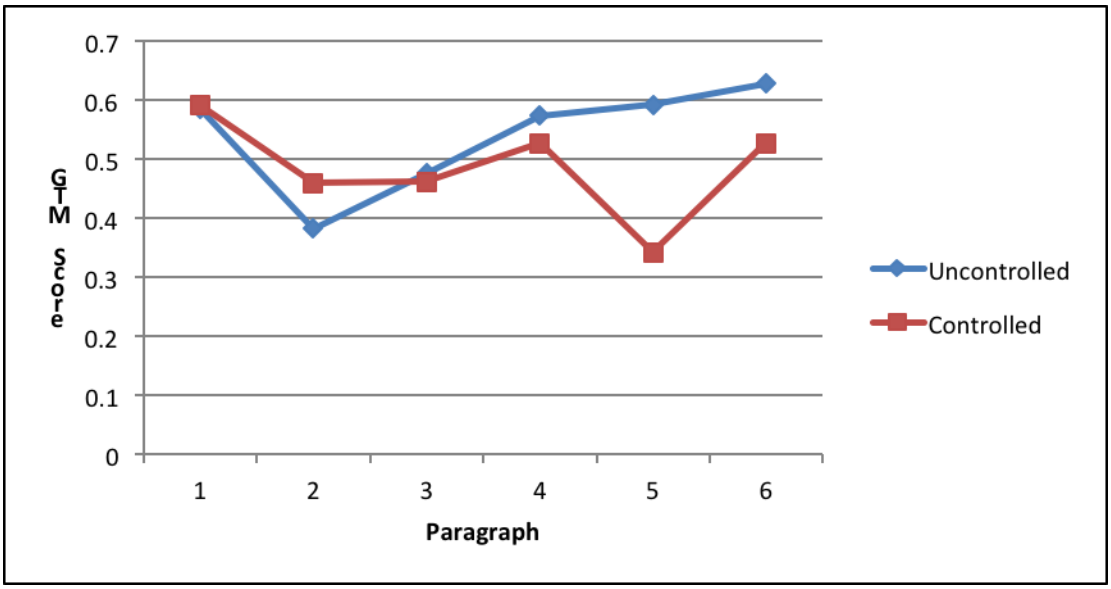

Figure 6.6: GTM Scores for Paragraphs for Both Conditions

6.2.4.3 BLEU

Higher BLEU scores also denote greater similarity to the reference translation. Table 6.6 shows the scores per paragraph and once again although the uncontrolled condition had a higher score (mean = 0.429 SD = 0.272) than the controlled (mean = 0.357, SD = 0.272), this difference was not significant as $df = 120$, $t = 1.471$, $p = 0.144$. Figure 6.7 gives these scores per paragraph.

Paragraph	1	2	3	4	5	6	Mean
Uncontrolled	0.456	0.179	0.358	0.507	0.549	0.591	0.429
Controlled	0.496	0.321	0.321	0.344	0.210	0.411	0.357

Table 6.6: BLEU Scores per Paragraph with Mean

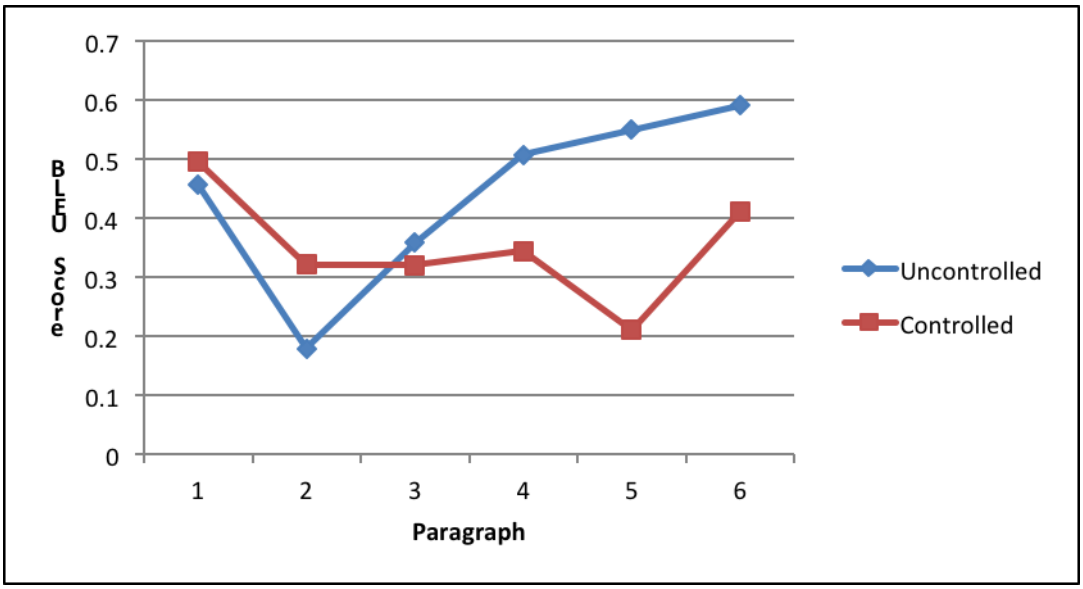

Figure 6.7: BLEU Scores for Paragraphs for Both Conditions

6.2.4.4 TER

As already indicated, the higher the TER score the more editing is required to meet the reference translation and therefore the 'worse' the MT is deemed to be. Table 6.7 shows these scores per paragraph. Once again the uncontrolled translation scored slightly better (mean = 0.498, SD = 0.257) than the controlled (mean = 0.537, SD = 0.248) and the difference was not significant where *df* = 118, *t* = -.847, *p* = 0.399. Figure 6.8 gives these scores per paragraph.

Paragraph	1	2	3	4	5	6	Mean
Uncontrolled	0.434	0.717	0.574	0.437	0.406	0.376	0.498
Controlled	0.404	0.566	0.539	0.493	0.684	0.523	0.537

Table 6.7: TER Scores per Paragraph with Mean

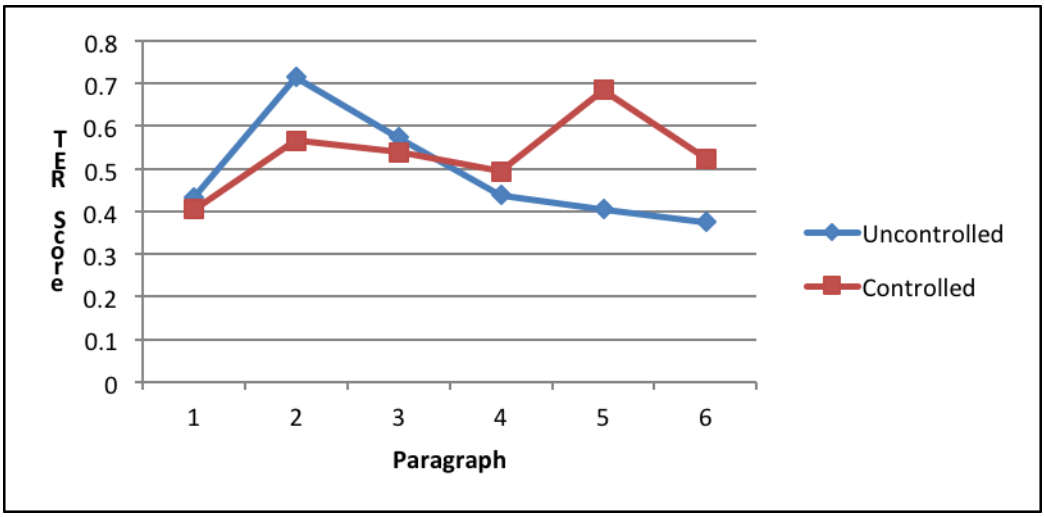

Figure 6.8: TER Scores for Paragraphs for Both Conditions

6.2.5 Within Grouping Correlational Analysis

Each of the AEMs was highly correlated with the others; Table 6.8 shows very high *r* scores that were all significant at the 0.01 level. This indicates a very strong agreement between each of the metrics and strengthens the individual findings presented above. Questions regarding validity of the findings would be raised if metrics measuring similar phenomena (e.g. text similarity using uni-grams) resulted in significantly different results. However, when scores from automatic metrics are consistent, it does not ensure that there are no differences between the translations in terms of their qualitative aspects, e.g. the use of synonyms.

	GTM	BLEU	TER
GTM	-	**.960****	**-.976****
BLEU	**.960****	-	**-.960****
TER	**-.976****	**-.960****	-

Table 6.8: Correlations between AEMs

When tested against Flesch and LIX (per paragraph), no significant correlations were found. Table 6.9 presents the *r* scores for each AEM and the other textual variables measurable per paragraph.

	Flesch	LIX
GTM	-.280 (.377)	.055 (.864)
BLEU	-.187 (.561)	.044 (.892)
TER	.276 (.385)	-.032 (.922)

Table 6.9: Correlations between AEMs and Per-Paragraph Measures

Due to the high level of co-variance between the AEMs, their correlation with other measures is very similar. Table 6.9 presents these values and suggests that readability as measured by the Flesch index does not correlate with translation quality as measured using edit-distance type AEMs. This is even more the case for LIX, given the very weak *r*-values for LIX and each of the AEMs. GTM shows slightly higher correlations with each of the per paragraph measures. However, given the overall range of largely weak *r* values it is evident that none

of the AEMS are strong indicators of readability and/or comprehensibility as measured by traditional readability indices of Flesch and LIX.

6.2.6 Discussion Points for Grouping A

Sample Size for Flesch/LIX

In the calculation of the t-test and correlation, the sample size of Flesch and LIX scores was problematic. This is because each *paragraph* was given a score due to the constraints of the formulae, namely that they cannot be used accurately on less than 100 words. Therefore the sample size, or number of observations/cases, was 6 given that 6 paragraphs were used in this study. The AEMs, on the other hand, produced a score for each sentence, which resulted in approximately 50 observations in each condition. Document/system-level scores were also generated by the AEMs to provide an overall system-level assessment of the quality of the MT output. When outliers were removed from the Flesch and LIX scores, the sample size was again made slightly smaller. Therefore, the ability to find a statistically significant difference was reduced as the likelihood of such a finding increases with sample size.

As described above, other studies have found no significant improvement in Flesch scores once controlled language has been implemented (e.g. Cadwell 2008, O'Brien 2010), and this is consistent with the findings here. However, given that sample size remains an issue in the current study, and possibly in other studies, it must be conceded that the validity of the t-test in testing for a difference between two conditions of such few observations is not an ideal way to test for a significant difference in the statistical sense. It is evident that there is a need for more observations (larger sample) for more accurate statistical power to overcome this shortcoming. Perhaps on a macro-level in, for example, an industrial scenario, where hundreds of paragraphs have been edited using *acrocheck* to implement the CL rule set, a significant improvement in Flesch scores would be observed. This is a topic that will be revisited in Chapter Seven.

Pilot Study and AEMs

Although the pilot study used only TER in its evaluation of individual sentences out of context, it is worth comparing the scores with those presented

in this section. The pilot found a moderate correlation between TER and the rated quality of sentences by human evaluators. TER was used as sentences were presented in isolation. In addition, it also showed a moderate correlation with observation length and fixation count both of which were significant ($p < 0.05$ in both cases). The relationship between the AEMs and the eye tracking metrics is dealt with later in this chapter.

Reference Translation

The most probable cause of the slightly better performance of the uncontrolled system is the higher similarity of its translations to its reference translations. As previously described, an in-house human translation (into French) was used as the reference when evaluating the MT output produced using the uncontrolled system. When the uncontrolled source text was edited to implement the CL rule set a new human translation had to be created to allow for a fair comparison. If this had not been done, then the uncontrolled system would have had an unfair advantage given its reference was based entirely on the uncontrolled source whereas the same reference would not wholly correspond to the content of the controlled source text. A new reference translation was provided by Symantec by an in-house translator with knowledge and expertise specific to the content. However, three pivotal variables may have had an effect, namely: a different translator produced the reference translation for the controlled text; the translation was carried out at least six years after the original translation; and the professional context/scenario may also have been different. While the change of translator and the time lapse do mean that these two new variables were introduced to the study, it was not possible to avoid this situation without biasing the experiment against the controlled system. However, this decision also presents potentially new biases.

6.2.7 Section Summary

This section began with a description of the evaluation of the MT output from both systems by means of error categorisation. It was found that the uncontrolled output had more errors than the controlled output and a slightly different distribution of errors across sub-categories. Secondly, the focus was moved to the traditional measures of readability, namely the Flesch and LIX indices. It was found that, although the controlled text resulted in better scores on both of these measures, the differences were not significant. However, given the small sample size, this was not surprising. There was a very strong negative correlation between Flesch and LIX, which indicates they both came to the same conclusion and supports their use in this context as a means of validly and accurately measuring readability via superficial linguistic phenomena. Following this, each of the three AEMs was examined. Overall, it was found that the uncontrolled text had slightly better results on all three measures: GTM, BLEU, and TER. However, this result was not found to be significant in any of the cases. There was a very strong correlation between all of the AEMs which again indicates their agreement in the findings. No significant correlation was found between the AEMs and the readability indices, which points to the fact that while one set of measures is examining textual similarity, the other is measuring linguistic complexity. Weak *r* values were found and suggest that none of the AEMS are strong indicators of readability as measured by traditional readability indices of Flesch and LIX. While each came to a valid within-measure conclusion, there is no correlation. Comparisons were then made to other studies in the respective areas and potential weaknesses of the current approach were highlighted. Specifically, small sample sizes in the calculation of Flesch and LIX scores were put forward as one explanation for the insignificant difference between the uncontrolled language and controlled language conditions. Finally, the problems caused by having to generate a new human reference translation for the controlled source text were also highlighted.

Grouping B: Eye Tracking Variables

6.3.1 Section Overview

This section consists of results and discussion of the findings from the eye tracking measures. The data are examined in the following order: observation length, fixation count, fixation length, pupil dilation, and measures of regression. On examination of the grouping as a whole, a MANOVA showed a significant difference between conditions where Pillai's $F = 1.346$, $df = 9.0$, $p < 0.05$, partial $\eta^2 = .548$. This result warrants a more detailed inspection of each variable and its interaction within this grouping, and also with the previous grouping of textual variables.

6.3.2 Observation Length

As described previously, observation length was defined as the duration in milliseconds for which the participant viewed each paragraph onscreen. While the pilot study, and other studies in the area, use the term 'Gaze Time', the software used in the main study replaced this term with 'Observation Length'. Table 6.10 provides the total observation length in seconds for all paragraphs (except paragraph 1), and the mean value per paragraph (i.e. total mean / five) for all participants (1 to 20) for both the uncontrolled and controlled conditions.

Uncontrolled	Total	Mean		Controlled	Total	Mean
1	423.820	84.764		**11**	211.020	42.204
2	299.726	59.945		**12**	475.080	95.016
3	253.014	50.603		**13**	433.267	86.653
4	465.916	93.183		**14**	413.182	82.636
5	319.214	63.843		**15**	323.100	64.620
6	295.637	59.127		**16**	279.676	55.935
7	392.218	78.444		**17**	229.031	45.806
8	370.002	74.000		**18**	385.199	77.040
9	314.768	62.954		**19**	355.507	71.101
10	265.397	53.079		**20**	221.914	44.383
Overall Mean	339.971	67.994			332.697	66.539

Table 6.10: Total and Mean Observation Length (seconds)

The initial impression from the data suggests that there is little difference between the conditions as their overall means are quite similar. An independent samples t-test found a non-significant difference between the conditions (t = 1.95, df = 18, p = .848), where the mean difference was 1.454 milliseconds. As the Table indicates, the uncontrolled condition resulted in a slightly greater mean for observation length (67.994 seconds, SD = 14.031, median = 63.398) than did the controlled condition (66.539 seconds, SD = 18.958, median = 67.860). The range of the uncontrolled condition was shorter: 42.58 with a minimum of 50.603 and maximum of 93.183; compared to the range of the controlled condition which was 52.812 with a minimum of 42.204 and a maximum of 95.016. Figure 6.9 shows the overall average observation length for each condition in seconds. It is evident that the uncontrolled condition resulted in longer average observation lengths for participants; however, as the t-test found, this difference is not statistically significant.

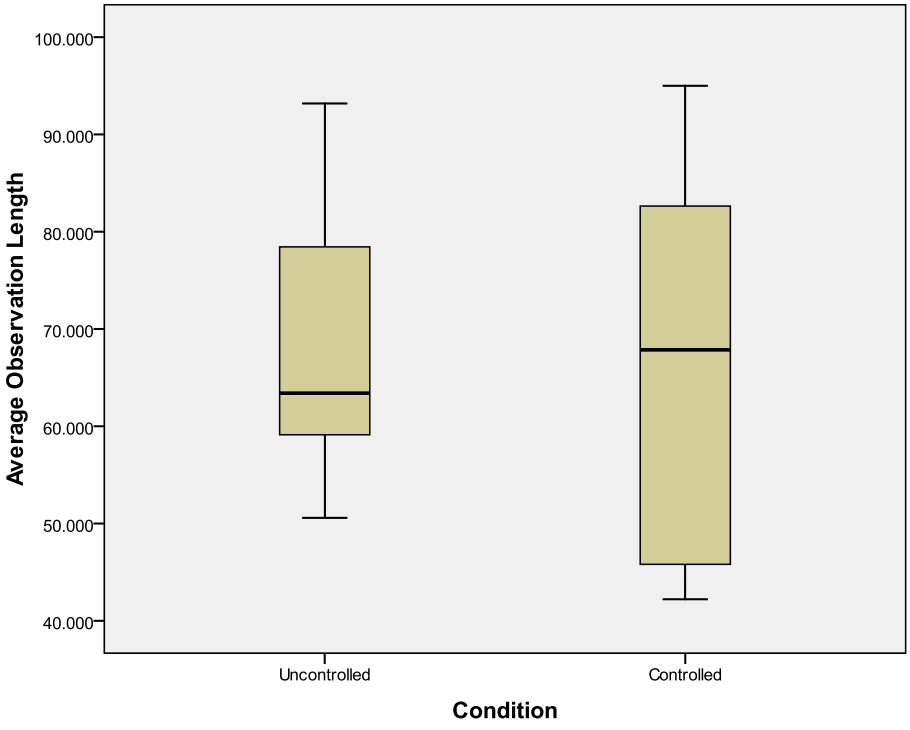

Figure 6.9: Average Observation Length for Both Conditions

These results can be contrasted with those of the pilot study, where observation length was found to correlate well with translation quality in that

sentences which were rated most favourably resulted in shorter observation lengths. Although there is no comparison of like with like here, as the conditions of the pilot study (translations judged to be 'good' versus translations judged to be 'bad') differ from the conditions to the current study (uncontrolled output versus controlled output), it is tempting to assume that the output from the controlled MT system is to translations judged 'good' as the output from the uncontrolled system is to the translations judged 'bad'. The results from the current study for average observation length suggest, however, that this analogy does not necessarily hold.

In a related study, Cadwell (2008) found no significant difference between texts of different Flesch scores in terms of reading speed. In contrast, O'Brien (2010) found that the controlled version of a text rated as easy to read resulted in longer observation length (than the uncontrolled version of the same text) and that the controlled version of a more difficult text took less time to read (than the uncontrolled version of the same text). The findings here add to such confounding results whereby the time taken by participants to read the uncontrolled version of the text was similar to the time taken for the controlled text.

6.3.3 Fixation Count

Fixation count, which is the number of fixations participants make onscreen, is summarised in Table 6.11. There appears to be a difference between conditions in that the controlled condition had fewer fixations overall and on average over the five paragraphs.

Uncontrolled	Total	Mean		Controlled	Total	Mean
1	1044	208		**11**	724	144
2	889	177		**12**	1372	274
3	596	119		**13**	760	152
4	1014	202		**14**	764	152
5	910	182		**15**	608	121
6	631	126		**16**	617	123
7	729	145		**17**	513	102
8	833	166		**18**	615	123
9	621	124		**19**	740	148
10	1249	249		**20**	495	99
Overall Mean	851	170			720	144

Table 6.11: Total and Mean Fixation Count

An independent samples t-test found a significant difference between the conditions on this measure, where t = 2.144, df = 18, p = .046 and a mean difference of 26.16. Further analysis shows that the uncontrolled condition had an overall mean of 851.60 (SD = 213.093, median = 861) fixations compared to the controlled's 720.8 (SD = 248.708, median = 670.50). The range of values in the uncontrolled condition was smaller at 653 (minimum of 596, maximum of 1249) against that of 877 (minimum of 495, maximum of 1372) for the controlled. Figure 6.10 shows these values for both conditions. As is evident from the box plot, an outlier (participant 12) can be identified. Upon removal from analysis, the independent t-test found a more significant difference between conditions where t = 2.593, df = 17, p = .019 with a mean difference of 40.631.

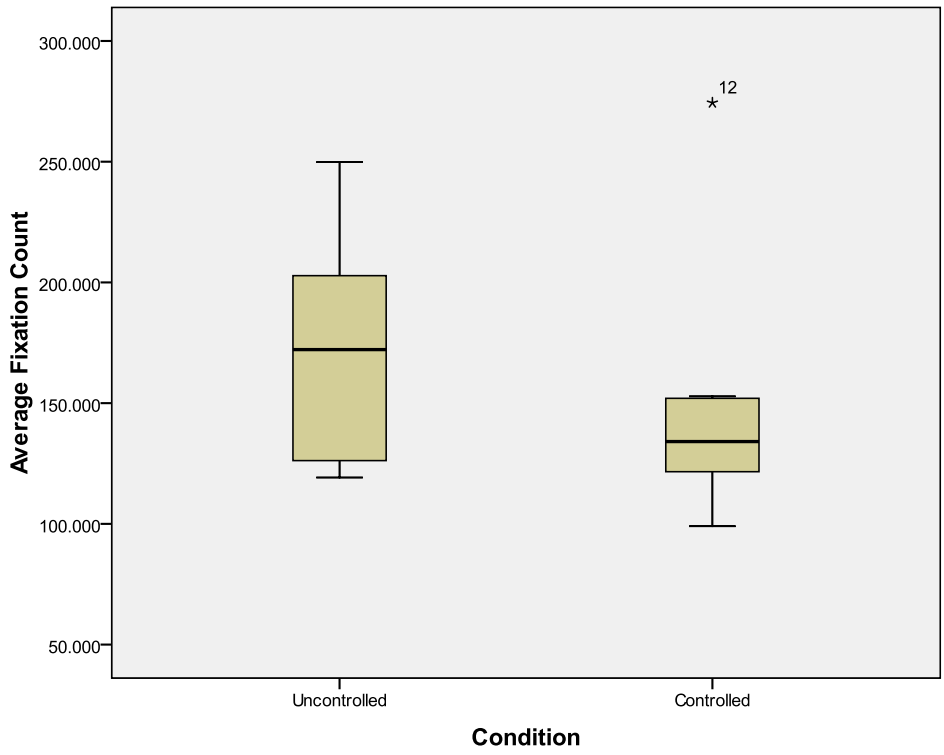

Figure 6.10: Average Fixation Count for Both Conditions

This result reflects the pilot study where fixation count was found to be a reliable indicator of translation quality in that higher rated sentences resulted in fewer fixations than those with lower ratings. In comparison with the pilot study's finding, and, if average fixation count can indeed be representative of cognitive effort, it can be construed that the controlled condition in the main study could be said to be of higher quality. The consistent finding between both studies on the measure of fixation count would support such an analogy. In addition, this result also indicates that while participants from both conditions spent a similar amount of time reading the text, the number of fixations was significantly higher in the uncontrolled condition.

In related studies, Sharmin *et al*. (2008) found that higher text complexity resulted in a higher number of fixations. However, O'Brien (2010) found that the controlled version of a 'difficult' text resulted in significantly fewer fixations, but on an 'easier' text this difference was not found. Such comparison must remain tentative as often in the literature terms such as quality and complexity are used in ways that are not comparable and, at times, these concepts are not operationalised or defined. Therefore, comparison between studies is not directly possible.

6.3.4 Fixation Length

Fixation length, which is the duration of fixations participants make onscreen, is summarised in Table 6.12. The Table shows that the controlled condition had shorter fixations overall and on average over the five paragraphs (both measured in seconds) than did the uncontrolled condition.

Uncontrolled	Total	Mean		Controlled	Total	Mean
1	187	37		**11**	127	25
2	155	31		**12**	132	26
3	101	20		**13**	122	24
4	146	29		**14**	125	25
5	143	28		**15**	120	24
6	153	30		**16**	109	21
7	110	22		**17**	107	21
8	123	24		**18**	115	23
9	129	25		**19**	139	27
10	124	24		**20**	102	20
Overall Mean	137	27			119	24

Table 6.12: Total and Mean Fixation Length for Both Conditions (seconds)

An independent samples t-test found a significant difference between the conditions on this measure where t = 2.194, df = 18, p = 0.042 with a mean difference of 3.773 seconds. Closer examination showed the fixations of the uncontrolled condition to be longer with a mean of 27.442 seconds (SD = 5.00, median = 27.24) compared to the controlled (mean = 23.669, SD = 2.134, median = 23.53). The range of the uncontrolled condition was also much larger: 17 (to the controlled's 7), where the minimum value was 20 (controlled = 20) and maximum 37 (controlled = 27). Figure 6.11 illustrates these values.

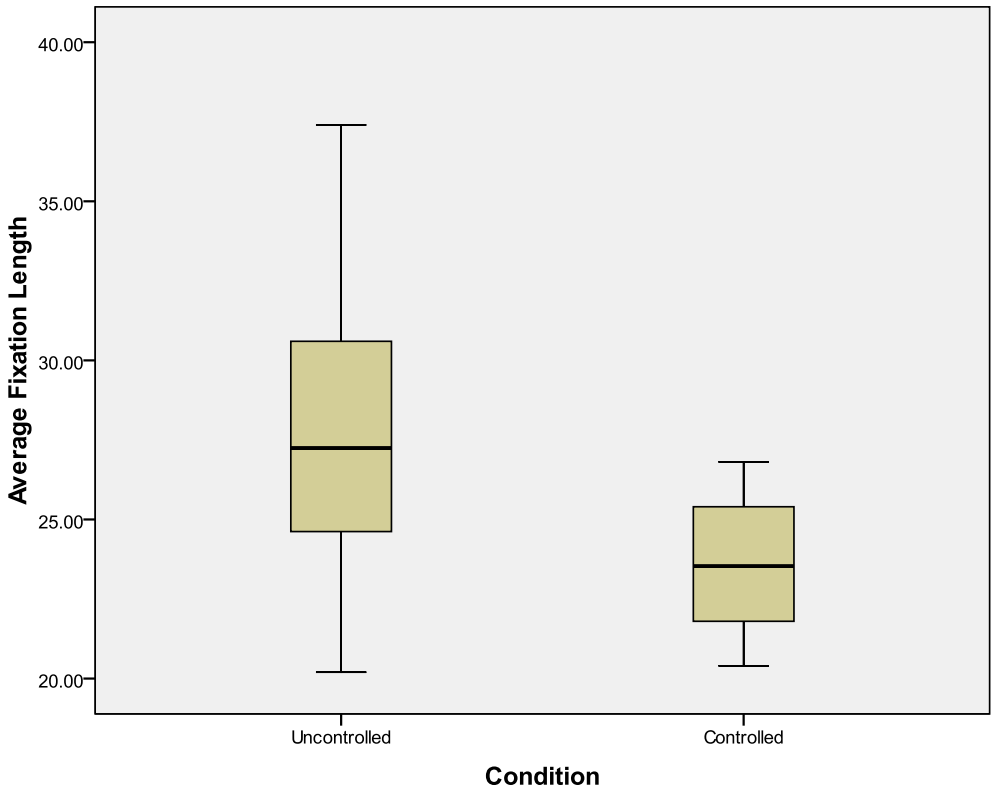

Figure 6.11: Average Fixation Length for Both Conditions

This finding contrasts with that of the pilot study where there was no significant difference between fixation length of the best and worst rated translated sentences, yet the highest rated sentences resulted in shorter fixations. It appears to be consistent, however, with the finding for fixation count in the current study. It also contrasts with recent studies such as O'Brien (2010), who concluded that fixation duration did not differ significantly between uncontrolled and controlled texts. In addition to this, Jakobsen and Jensen (2008) also found insignificant differences in fixation lengths across groups in translation processes as did Sharmin *et al.* (2008). Although the experimental conditions are once again different, the conceptual framework allows for an interesting comparison.

6.3.5 Percentage Change in Pupil Dilation (PCPD)

By measuring the diameter of the pupil during the task, its dilation and contraction can be observed from the onset of the stimulus of each paragraph

until the participant moves on to the blank white slide prior to the onset of the next paragraph. As indicated in Chapter Five, a latency effect of 1500ms was compensated for in the current study and median values were used in analysing the raw output where pupil size is measured in millimetres. The median values were used to calculate a baseline across the entire task due to the reported individual variance in pupil sizes (O'Brien 2010). The baseline is unique to each participant and attempts to represent the standard size of the pupils (one value is given by using the average size of both pupils) of the participant during the reading/evaluation task so that changes can be compared against these values. The values below identify percentage decreases and increases from the baseline value. For example, participant 1's pupil size decreased, on average, by 25% across the task compared to the baseline. Once again, all values exclude the first paragraph. Table 6.13 provides an overview of the PCPD for each conditions and their overall means (in percentages).

Uncontrolled	PCPD		Controlled	PCPD
1	-25		**11**	-19
2	-36		**12**	-37
3	-36		**13**	-38
4	-1		**14**	-9
5	-63		**15**	+14
6	+8		**16**	-39
7	-62		**17**	-50
8	-58		**18**	+9
9	+20		**19**	-39
10	-48		**20**	-49
Overall Mean	-30			-26

Table 6.13: Overall Values for PCPD in Percentages

There was little difference between the conditions in terms of their overall PCPD values. However, the uncontrolled condition was slightly larger by 35%. An independent samples t-test found this difference not to be of significance: t = -.259, df = 18, p = .47. The uncontrolled condition had a mean of -601 (SD = 3, median = -36), while the controlled condition had a mean of -259 (SD = 51, median = -38). The range for the uncontrolled condition was far less (range = 57, minimum = -63, maximum = 20) than the controlled (range = 88, minimum = -50, maximum = 38.). Figure 6.12 illustrates these values.

Upon inspection, the presence of two outliers (participants 14 and 18) in the controlled condition can explain the large range and standard deviation. Once removed, an independent samples t-test found the difference to remain insignificant, yet closer to the threshold: $t = .162$, $df = 16$, $p = .182$.

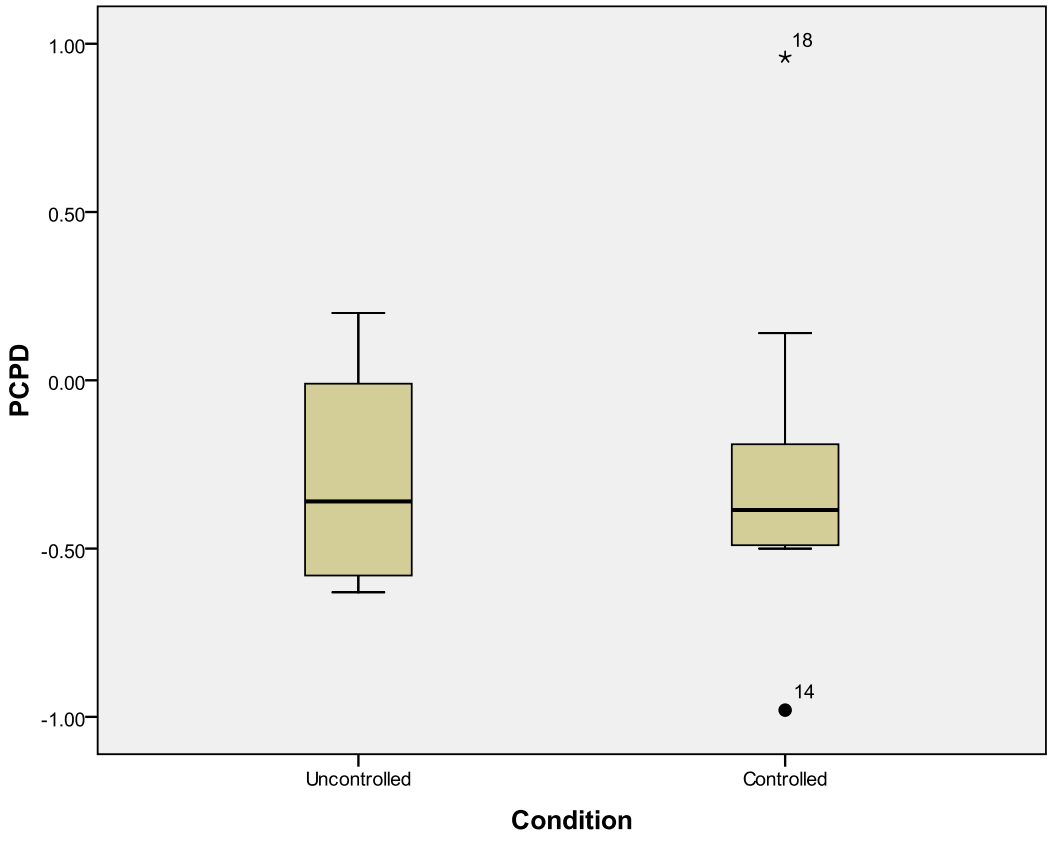

Figure 6.12: PCPD for Both Conditions

These findings are similar to those of the pilot study which found little difference in pupil dilation between sentences rated to be of high and low quality translation. Indeed, pupil dilation has recently be found to not correlate well with other eye tracking measures in similar studies (Schultheis and Jameson 2004, O'Brien 2008), yet its use as an indicator of cognitive effort and task difficulty is well documented (Rayner 1998). It would appear that once again the use of pupil dilation in this context has been questionable. Although the controlled text (which one might assume to be the 'easier' text) resulted in less of a change in PCPD overall, the difference was not significant. It would also indicate that, as described in the pilot study, pupil dilation does not correlate

well with other eye tracking measures and represents perhaps a different aspect of cognitive load (Iqbal *et al*. 2005) or may not be suitable for measuring readability and comprehensibility in this scenario.

6.3.6 Regressions

As described previously, regressions were counted manually via the Tobii Studio replay function. The total number of regressions for the five paragraphs was counted for each participant and an overall mean per participant was calculated in each condition. Individual means per paragraph were also calculated for each participant and an overall paragraph mean calculated for each condition. Table 6.14 provides these data.

Uncontrolled	Total	Mean		Controlled	Total	Mean
1	223	44		**11**	105	21
2	106	21		**12**	223	44
3	176	35		**13**	89	17
4	234	46		**14**	96	19
5	142	28		**15**	204	40
6	254	50		**16**	241	48
7	202	40		**17**	168	33
8	203	40		**18**	193	38
9	241	48		**19**	155	31
10	231	46		**20**	86	17
Overall Mean	201	40			156	31

Table 6.14: Total and Mean Number of Regressions

Table 6.14 shows that the uncontrolled condition had, on average, a higher number of regressions (mean = 201.2, SD = 47.16, median = 213) than the controlled condition (mean = 156, SD = 58.8, median 161.5). The range for the uncontrolled condition was, however, smaller (range = 148, minimum = 106, maximum = 254) than that of the controlled condition (range = 155, minimum = 86, maximum = 241). Figure 6.13 shows the average number of regressions for both conditions per paragraph. An independent samples t-test found this difference to be close to, but not significant where $t = 1.886$, $df = 17$, $p = .071$ with a mean difference of 45.32 regressions (outlier of participant 2 removed).

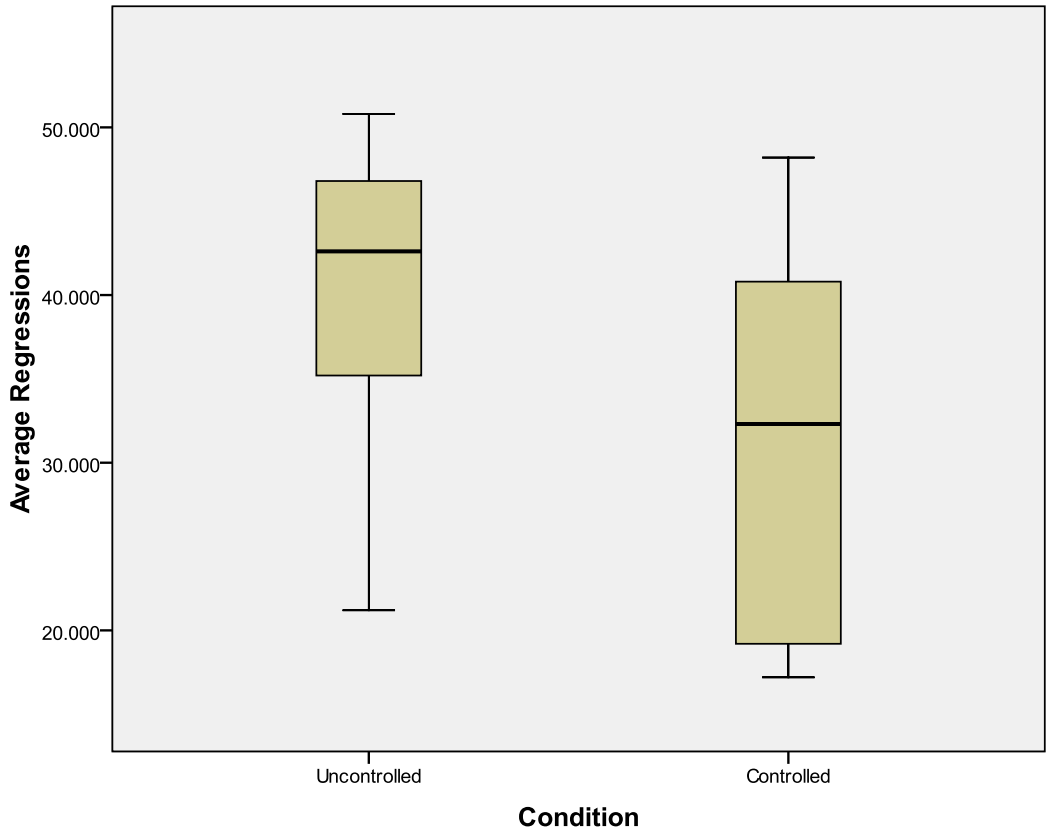

Figure 6.13: Average Number of Regressions per Paragraph

6.3.7 Regression Distance

To provide more detail on the nature of these regressions, each was analysed in terms of its distance in words (Rayner 1998). The uncontrolled condition resulted in a mean of 3.06 words per regression (SD = .833, median = 2.99) whereas the controlled condition had a lower value of 2.93 words per regression (SD = .860, median = 2.91). Ranges for both conditions were similar where the uncontrolled had a value of 2.42 (minimum = 1.95, maximum = 4.37) and the controlled 2.18 (minimum = 1.97, maximum = 4.15). Table 6.15 and Figure 6.14 illustrate these figures.

	P2	P3	P4	P5	P6
Uncontrolled	2.94	2.94	2.88	2.86	3.04
Controlled	2.94	2.87	3.14	3.03	3.30

Table 6.15: Mean Values per Paragraph for Regression Distance in Words

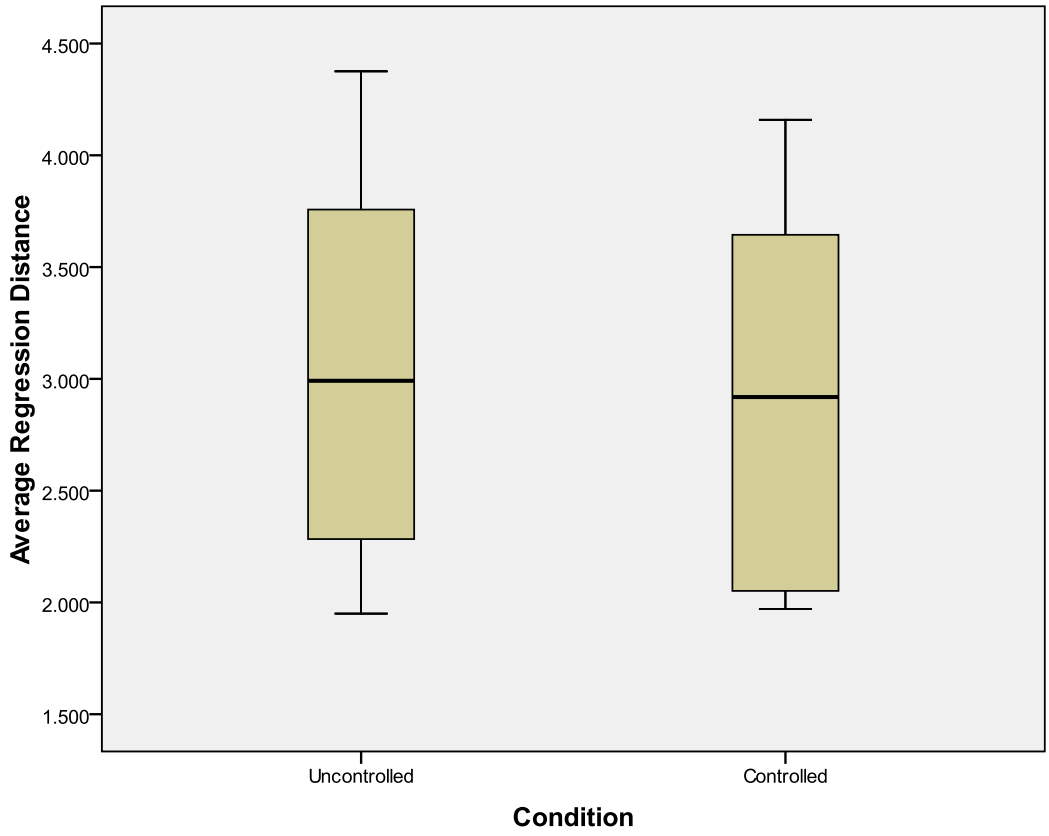

Figure 6.14: Average Regression Distance in Words

An independent samples t-test found no significant difference between conditions where $t = .324$, $df = 18$, $p = .749$. During the analysis of this measure, it became evident that the controlled condition contained a higher number of one-word regressions and the uncontrolled condition was higher in regressions where regression distance was greater than ten words which points to comprehension failure (Rayner 1998). In this respect, it is also worth noting the median values for regressions per paragraph to ensure that the mean is not distorted by these clusters of frequent one-word regressions and indeed by the higher values. Table 6.16 provides these data. Using the median values, an independent samples t-test found the difference between conditions to be highly significant where $t = 3.66$, $df = 18$, $p = .002$.

Paragraph	**2**	**3**	**4**	**5**	**6**
Uncontrolled	1.8	1.4	2.5	2.2	1.8
Controlled	1.3	1.4	1.4	1.35	1.5

Table 6.16: Median Values per Paragraph for Regression Distance in Words

Lastly, Table 6.17 shows that the number of regressions equal to or greater than ten words was found to be higher in the uncontrolled condition (mean = 21, SD = 10.381, median = 20) than the controlled (mean = 13, SD = 8.954, median = 14).

Condition	Distance ≥ 10
Uncontrolled	21
Controlled	13.3

Table 6.17: Regression Distance ≥ 10 Words

The range is somewhat larger for the uncontrolled condition at 33 (minimum = 11, maximum = 44) than the controlled condition's 23 (minimum = 0, maximum = 23). A highly significant difference was found where $t = 2.898$, $df = 18$, $p = .01$. Figure 6.15 illustrates the values for both conditions. The presence of an outlier (participant 3) warrants its removal. Upon reanalysis, the difference remains significant where $t = 2.664$, $df = 17$, $p = .016$.

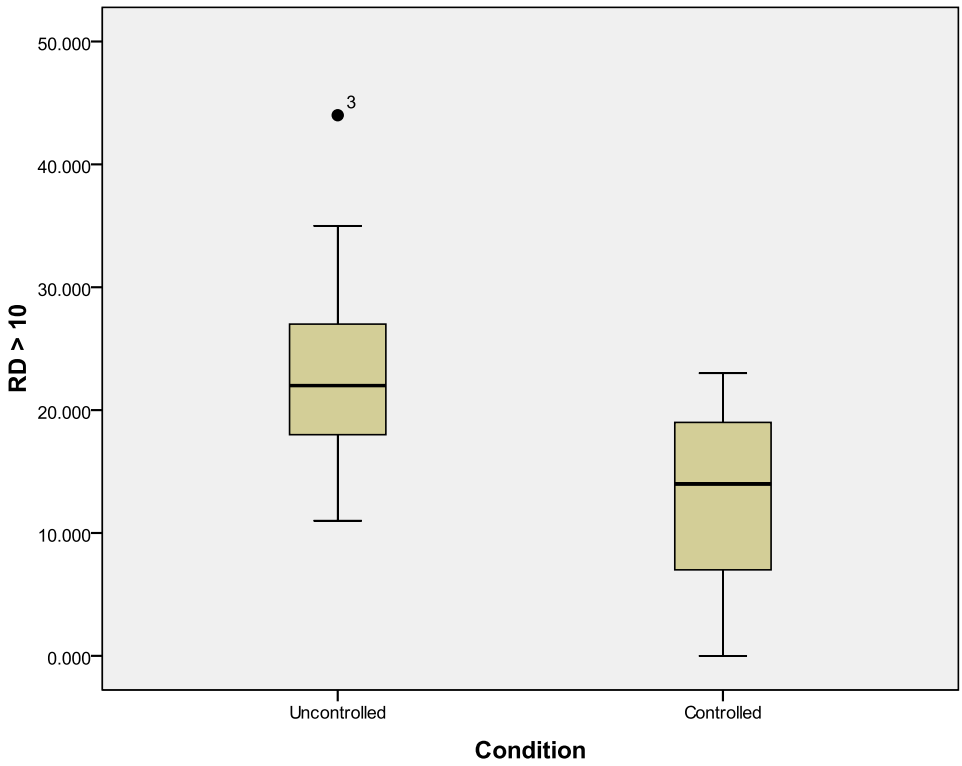

Figure 6.15: Regression Distance ≥ 10 Words for Both Conditions

6.3.8 Within Grouping Correlational Analysis

An analysis of the relationships between the eye tracking variables with one another provides a more comprehensive view of the results, and supplements the findings in the previous sections. This analysis used the scores for each eye tracking variable as measured per participant from each condition. An overview of the correlation coefficients (and *p* values) can be found in Table 6.18. As a reminder: significant correlations are highlighted in bold, one asterisk denotes significance at the 0.05 level, and two denote significance at the 0.01 level.

	Obs. Length	Fix. Count	Fix. Length	PCPD	Reg.	Mean Distance	Median Distance	Distance ≥ 10
Obs. Length	-	**.517* (.02)**	.354 (.125)	-.037 (.875)	.194 (.412)	-.216 (.359)	.077 (.749)	.053 (.823)
Fix. Count	**.517* (.02)**	-	**.498* (.025)**	-.186 (.433)	-.117 (.622)	.157 (.509)	.234 (.322)	.247 (.293)
Fix. Length	.354 (.125)	**.498* (.025)**	-	-.066 (.784)	.078 (.743)	.232 (.325)	.406 (.076)	.228 (.334)
PCPD	-.037 (.875)	-.186 (.433)	-.066 (.784)	-	-.276 (.239)	-.185 (.436)	-.331 (.154)	.141 (.552)
Reg.	.194 (.412)	-.117 (.622)	.078 (.743)	-.276 (.239)	-	.009 (.968)	**.546** (.01)**	**.43* (.05)**
Mean Distance	-.216 (.359)	.157 (.509)	.232 (.325)	-.185 (.436)	.009 (.968)	-	.119 (.617)	**.442* (.05)**
Median Distance	.077 (.749)	.234 (.322)	.406 (.076)	-.331 (.154)	**.546** (.01)**	.119 (.617)	-	.349 (.131)
Distance ≥ 10	.053 (.823)	.247 (.293)	.228 (.334)	.141 (.552)	**.43* (.05)**	**.442* (.05)**	.349 (.131)	-

Table 6.18: Correlations in Grouping B

With regard to the main eye tracking metrics, observation length was found to have a moderate correlation with fixation count (r = .517, p = .02), but not with fixation length, PCPD, or any of the regression measures in this grouping. In addition to its correlation with observation length, fixation count was also correlated with fixation length (r = .498, p = .025); however, no significant correlation between it and the other measures was found. PCPD did not correlate significantly with any other eye tracking measure.

Concerning the local correlational analysis of the regression variables, the mean number of regressions does not correlate with the mean distance of the

regressions where r = .009. This is a very weak value and indicates a very random and non-linear relationship. However, the mean number of regressions does show a moderate correlation with the median values for regression distance where r = .546 (highly significant at the p = .01 level), and with the number of regressions equal to or greater than ten words (r = .43, p = .05). These correlations indicate that the use of the median value in the measurement of regression distance is more in line with the number of regressions overall.

In sum, these findings indicate that there is an insignificant relationship between the number of fixations and the distance travelled, but the analysis highlights an interesting dimension to the examination of regressions, i.e. taking the distance dimension into account. Furthermore, the number of regressions equal to or greater than ten words shows a strong correlation with the total number of regressions and their mean distance values. This indicates that the greater the number of regressions, the more likely it is very long regressions will be present. Lastly, median regression distance appears to be a more accurate representation of regression distance than mean regression distance.

6.3.9 Analysis of Interaction between Grouping A and B

On examination of the relationships between the eye tracking variables in this grouping and the readability indices of grouping A, some interesting findings are evident; Table 6.19 provides an overview of the *r-values* with the significance level in parenthesis.

	Obs. Length	Fix. Count	Fix. Length	PCPD	Regression	Mean Distance	Median Distance
Flesch	**-.651* (.042)**	.466 (.175)	.193 (.594)	**.542* (.05)**	-.113 (.756)	.079 (.828)	.343 (.333)
LIX	.457 (.158)	**-.693* (.018)**	-.058 (.865)	**-.725* (.012)**	-.026 (.938)	.179 (.598)	-.194 (.598)
GTM	.427 (.166)	-.083 (.797)	.202 (.528)	-.079 (.806)	-0.29 (.928)	.180 (.576)	.242 (.448)
BLEU	.124 (.701)	.064 (.843)	.029 (.928)	.182 (.572)	-.433 (.16)	.345 (.272)	.491 (.105)
TER	.428 (.165)	-.084 (.795)	.203 (.527)	-.081 (.803)	-.027 (.934)	.179 (.578)	.24 (.451)

Table 6.19: Correlations between Grouping A and B

First of all, observation length showed a significantly moderate negative correlation with Flesch (r = -.651, p = .042), but not with LIX (r = .457, p =.158). Secondly, fixation count had no significant correlation with Flesch (r = .466, p = .175), but one was found with LIX (r =-.693, p = .018). Thirdly, fixation length also had no significant correlations with either measure: for Flesch (r = .193, p = .594 and for LIX (r = -.058, p = .865). Fourthly, PCPD was strongly correlated with LIX (r = -.725, p = .012) and moderately correlated with Flesch (r = .542, p = .05).

With regard to the measures of regression: the total/average number of regressions had no significant correlation with either Flesch (r = -.113, p = .756) or LIX (r = -.026, p = .938), nor did regression distance using the mean (where Flesch had an r value = .079, p = .828, and LIX r = .179, p = .598) or using the median (Flesch had an r value = .343, p = .333 and LIX had and r value = -.194, p = .598). As Table 7.18 shows: no significant correlations were found between the eye tracking measures and the AEMs. BLEU demonstrated a moderate correlation with median regression distance; however, it was not significant. The

other correlations had otherwise quite weak *r*-values. In sum, these results indicate Flesch had significant correlations with observation length, and with PCPD, while LIX correlated with PCPD, and with fixation count.

6.3.10 Discussion Points for Grouping B

Pilot Study

The findings presented in this section provide several contrasts with the pilot study, and indeed other studies in the area. They also highlight the problems posed in using eye tracking where such measures as fixations and observation lengths do not always yield results consistent with other studies, a point that will be revisited in Chapter Seven. The differences in the findings between the measures themselves is also concerning in that such contradictory results point to confounding factors and question the validity of the method in this capacity.

Inconsistency of Eye Tracking Measures

It is evident from the pilot study and the current study that there are inconsistencies between results gained using different eye tracking metrics. In other words, measures that correlated in the pilot study, for example, observation length and fixation count, do not correlate significantly in the main study. Although the experimental design of the studies differed, it was expected that such agreements would be repeated. This highlights a need for further testing and refinement of the methodology, or perhaps reflects shortcomings in the theoretical framework of the study, especially with regard to the cognitive process assumed to cause physiological changes in the characteristics and behaviour of the eyes. As highlighted in the review of literature, inconsistency or only using one or two metrics in a study, are common occurrences in other studies employing eye tracking methodologies, and this fact underscores the need for other points of data collection in a study, e.g. via human evaluation and recall tests.

6.3.11 Section Summary

This section examined the second grouping which was composed of the eye tracking measures. Each measure was examined in detail and then an overall within-grouping analysis was conducted followed by a comparison vis-à-vis the textual variables from Grouping A.

Firstly, the analysis of variance between the uncontrolled and controlled conditions found no significant difference in terms of observation length or PCPD. Both fixation count and fixation length were significantly different between conditions, where the controlled condition had, on average, fewer fixations and shorter fixation lengths. With regard to the measurement of regressions, the mean number of regressions did not differ significantly between conditions, nor did mean regression distance. A significant difference was found between conditions, however, in terms of median regression distance and the number of regressions of distance ≥ 10 words.

Secondly, a correlation analysis found that observation length correlated well with fixation count but no other eye tracking measures. Fixation count was also correlated strongly with fixation length, which in turn, correlated with GTS, while PCPD had no correlates within this grouping. For the measures of regression, the number of regressions showed a significant moderate correlation with median regression distance, and with the number of regressions of distance ≥ 10. This finding indicates that the median value in the measurement of regression distance is more in line with the number of regressions.

Finally, an examination of the eye tracking variables with the textual variables from Grouping A found that observation length had a significant correlation with the Flesch readability index, and that both LIX and Flesch were correlates of PCPD. Therefore, it can be posited that Flesch is a predictor (see section 6.5) of reading time, i.e. observation length, yet the insignificant correlation with LIX highlights the difference in the two readability indices. As described previously, Flesch uses the number of syllables, words, and sentences in its calculation, while LIX uses the number of words, full stops (periods), and words greater than 6 letters. It is probable that such differences in the

calculation of the scores between the indices accounts for an effect observed in observation length, and are worthy of further investigation.

Grouping C: Human Evaluation Variables

6.4.1 Section Overview

The third grouping of human evaluation variables draws on the user-based evaluation of MT output. As previously described, participants were provided with a hard-copy of the respective paragraphs from their condition, and asked to evaluate them in terms of their readability and comprehensibility. Operationalised definitions of readability and comprehensibility were given beforehand and also appeared on the bottom of each page to serve as a reminder. After this was completed, participants were asked to take a recall test to examine how much of the content from the paragraphs they could remember. As comprehensibility and recall are conceptually linked in this study, they will be examined together after the results and discussion of the readability evaluation scores.

Overall, a MANOVA showed a significant difference between conditions for the above variables where Pillai's $F = 4.291$, $df = 3.0$, $p < 0.05$, partial $\eta^2 = .446$; therefore closer inspection of each variable is necessary. It should be noted that both measures of readability and comprehensibility showed a high level of internal reliability with a Cronbach's α of .741 and .824 respectively (a value of .7 and above is satisfactory) which supports their use in this context. The reader is reminded here that these values represent percentages as each condition had a different number of sentences; therefore the scores had to be normalised (*z*-scores) and percentages were used for a clearer description.

6.4.2 Readability

As previously detailed, the evaluation of the texts was carried out by the participants, who were asked to rate the sentences in terms of their readability. This provides an interesting comparison to the objective readability indices (Flesch and LIX) discussed earlier. Table 6.20 provides an overview of the scores given by participants to the individual in-context sentences in their condition.

Uncontrolled	Mean		Controlled	Mean
1	65.36		**11**	69.03
2	51.84		**12**	73.11
3	49.90		**13**	57.19
4	58.58		**14**	71.88
5	70.43		**15**	68.45
6	52.64		**16**	67.49
7	69.14		**17**	69.83
8	70.75		**18**	67.78
9	75.48		**19**	67.83
10	63.37		**20**	65.84
Overall Mean	62.75			67.84

Table 6.20: Mean Readability Evaluation Scores for Both Conditions in Percentages

The uncontrolled condition had a mean of 62.75% (SD = 9.04), a median of 64.37, and a range of 25.58 (min. = 49.9, max. = 75.48), and the controlled condition had a greater mean of 67.84% (SD = 4.31), a median of 68.14, and a range of 15.91 (min. = 57.19, max. = 73.11). An independent samples t-test found the difference between conditions not to be significant where t = -1.607, df = 18, p = .12. These values are indicated in Figure 6.16, which also highlights the presence of an outlier (participant 13). Upon removal, the difference is close to, but not, significant where t = -2.019, df = 17, p = 0.059.

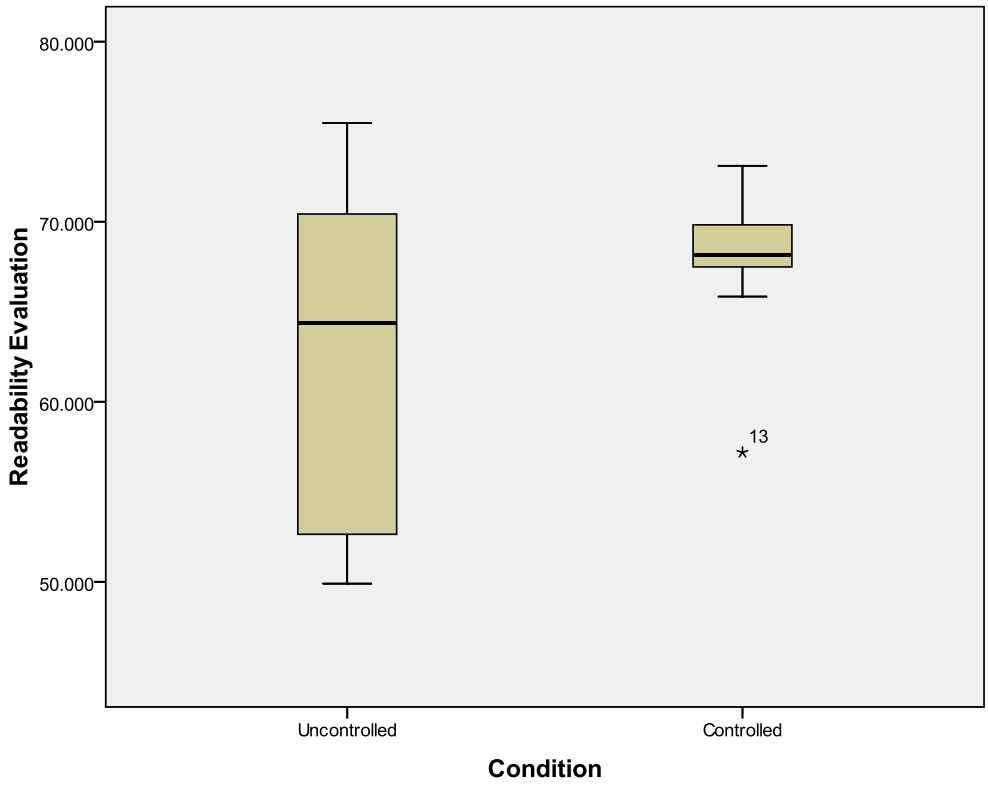

Figure 6.16: Average Readability Scores for Both Conditions

As readability was not measured in this way in the pilot study, no directly comparable results exist. O'Brien (2010) found that controlled texts did not result in better ratings from human evaluators in terms of perceived readability. Yet Cadwell (2008) found the majority of participants rated the controlled texts in his experiment "easier to read" (ibid., p. 40). However, Bernth and Gdaniec (2001) showed that after applying MT-oriented CL rules readability of the controlled text used in their experiment was, in fact, reduced.

6.4.3 Comprehensibility

The results from the comprehensibility evaluation show a similar trend to the readability scores. Table 6.21 provides the scores and shows that the controlled condition had a higher rating for comprehensibility.

Uncontrolled	Mean		Controlled	Mean
1	81.31		**11**	64.66
2	65.61		**12**	83.68
3	49.44		**13**	66.35
4	60.22		**14**	76.81
5	66.99		**15**	82.77
6	51.21		**16**	77.9
7	70.60		**17**	78.78
8	77.00		**18**	80.60
9	77.44		**19**	70.65
10	68.66		**20**	66.73
Overall Mean	66.85			74.89

Table 6.21: Mean Comprehensibility Evaluation Scores for Both Conditions in Percentages

As shown in Table 6.21 and Figure 6.17, the uncontrolled condition had a mean of 66.85% (SD = 10.72), a median of 67.82, and a range of 31.87 (min. = 49.44, max. = 81.31), and the controlled condition had a higher mean of 74.9 (SD = 7.16), a median of 77.38, and a range of 19.01 (min. = 64.66, max. = 83.68). An independent samples t-test showed that the difference between the conditions was once again, close to, but not statistically significant (t = -1.974, df = 18, p = 0.054).

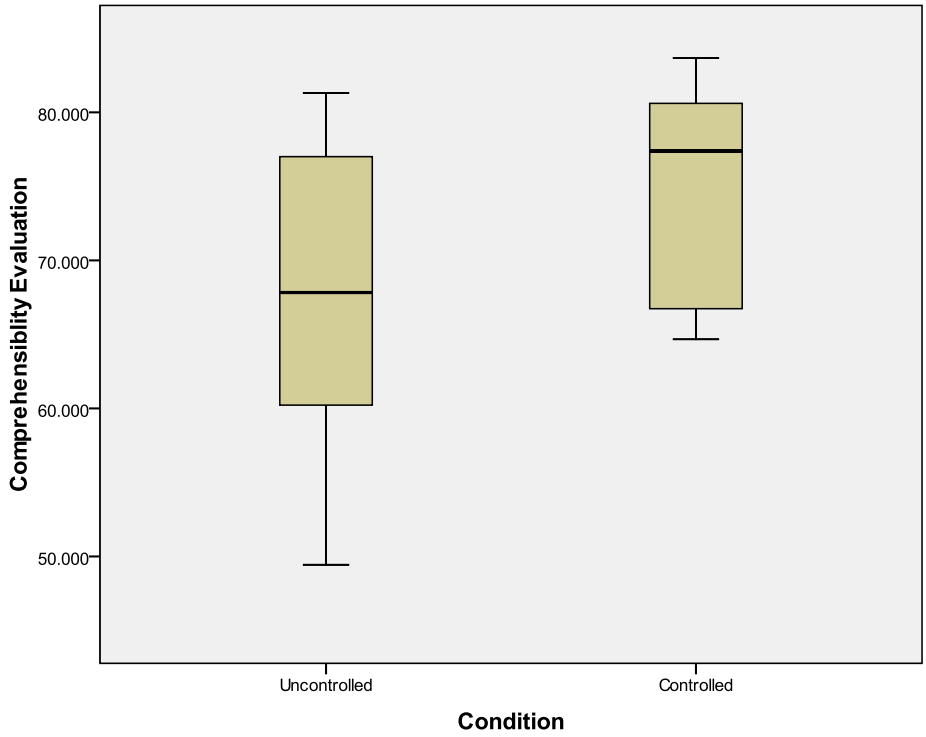

Figure 6.17: Average Comprehensibility Scores for Both Conditions

Interestingly, in both conditions, all but two of the participants rated the texts overall as being more comprehensible than readable, or in other words, scores for comprehensibility were higher than readability in all but two cases. This may indicate the participants' ability (or perceived ability) to understand the content despite readability difficulties. Readability and comprehensibility showed a strong positive correlation where $r = 0.848$ ($p < 0.001$), thus indicating their co-variance and construct validity as evident in the linear relationship shown in Figure 6.18. In other words, as readability increased so too did comprehension.

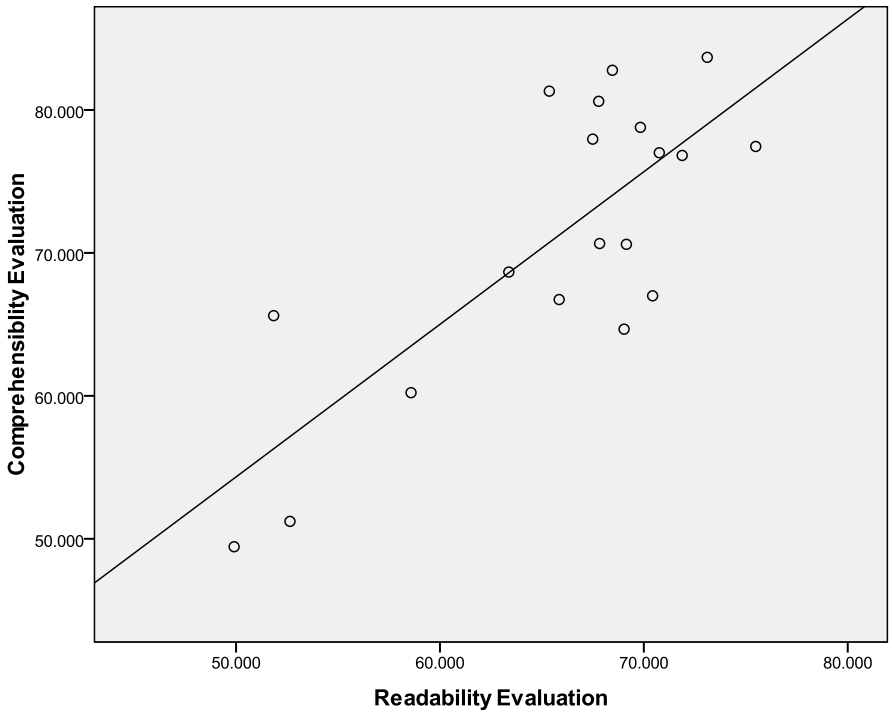

Figure 6.18: Correlation between Readability and Comprehension Evaluation Scores

6.4.4 Recall Test

As described in earlier chapters, participants were tested for recall on three levels by means of general questioning, cloze testing, and cued recall. The recall test has a high level of internal reliability with a Cronbach's α of 0.707 (values of .7 and above are satisfactory). Table 6.22 provides an overview of the total number of items recalled out of a possible score of 20. The relatively low number of items recalled is unsurprising - DuBay (2004) states that all readers, even those who are deemed to be advanced, have a limitation of 65% on the amount they can recall from a simple text. Given the novelty of the experiment to the participant and their acknowledged inexperience with the textual genre, values under this were expected.

Uncontrolled	Mean		Controlled	Mean
1	16		**11**	11
2	4		**12**	17
3	0		**13**	12
4	5		**14**	7
5	3		**15**	11
6	4		**16**	6
7	0		**17**	10
8	3		**18**	10
9	1		**19**	8
10	6		**20**	9
Overall Mean	4.2			10.1

Table 6.22: Total Recall Scores for Both Conditions

The uncontrolled condition recalled fewer items with a mean of 4.2 (SD = 4.6, median 3.5) to the controlled condition's 10.1 (SD = 3.07, median 10). Also, the range of the uncontrolled condition was wider with a value of 16 (min. = 0, max. = 16) compared to a value of 11 (min. = 6, max. = 17) for the controlled condition. Of interest is that each participant in the controlled condition recalled at least some items, while two participants in the uncontrolled condition recalled nothing (as measured in this context). An independent samples t-test found a very significant difference between the two conditions (t = -3.366, df = 18, p = 0.003). Figure 6.19 illustrates these data and highlights participants 1 and 12 as outliers. Upon removal, an independent samples t-test found a significant

difference between the two conditions with higher t and lower p values (t = -6.588, df = 16, p < 0.001).

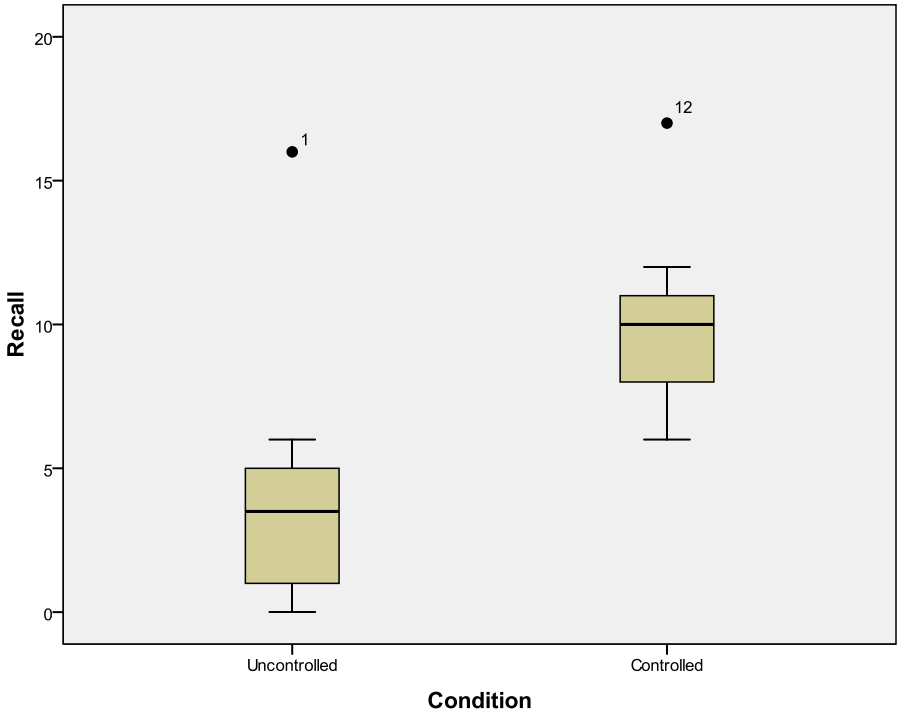

Figure 6.19: Total Recall Scores for Both Conditions

Lastly, the human evaluation of comprehensibility correlates well with items recalled (r = .511, p = .021) which supports the operationalisation of comprehensibility in this context. Figure 6.20 shows this relationship, and Table 6.23 provides the correlation coefficients within this grouping.

	Readability	**Comprehensibility**	**Recall**
Readability	-	**.848**** **(.001)**	.385 (.127)
Comprehensibility	**.848**** **(.001)**	-	**.511*** **(.021)**
Recall	.385 (.127)	**.511*** **(.021)**	-

Table 6.23: Correlations in Grouping C

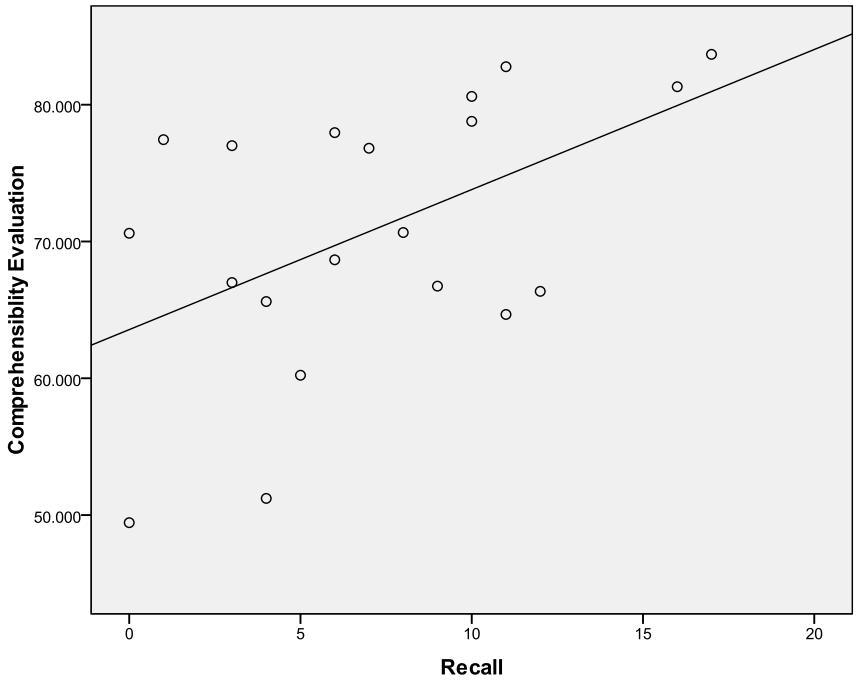

Figure 6.20: Correlation between Recall and Comprehensibility Evaluation Scores

In similar work, O'Brien (2010) found a slightly higher level of recall for controlled texts, but this effect was somewhat negated by items incorrectly recalled. Cadwell (2008) also found recall was slightly higher (in lists) for controlled texts. However, the recall tests used in these studies differ from the test employed here.

6.4.5 Analysis of Interaction between Groupings

A summary of the correlations between the textual variables of Grouping A and the human evaluation variables of Grouping C can be found in Table 6.24. No significant correlations were found between either of the human evaluation scores for readability and comprehensibility and the textual variables, however, recall was found to be significantly correlated with Flesch scores (r = .529, p = .041).

	Flesch	LIX	GTM	BLEU	TER
Readability	.287 (.422)	-.312 (.35)	.279 (.407)	.126 (.696)	-.293 (.382)
Comprehensibility	.079 (.827)	.086 (.801)	.37 (.263)	.109 (.735)	-.38 (.249)
Recall	**.529* (.041)**	-.136 (.726)	-.279 (.467)	-.35 (.322)	.364 (.362)

Figure 6.24: Correlations between Groupings C and A

A summary of the correlations between the eye tracking variables of Grouping B and the human evaluation variables of Grouping C can be found in Table 6.25. A significant moderate negative correlation was found between readability and regression distance where r = -.544 (p = .016). Comprehensibility had no significant or moderate/strong correlations with any of the eye tracking variables. However, recall had significant moderate correlations with both PCPD (r = .514, p = .05) and regressions (r = -.475, p = .046).

	Obs. Length	Fix. Count	Fix. Length	PCPD	Reg.	Reg. Distance	Reg. Distance ≥ 10
Readability	.22 (.366)	-.136 (.591)	-.184 (.451)	-.213 (.428)	-.022 (.928)	**-.544* (.016)**	-.332 (.178)
Comprehensibility	.21 (.375)	-.076 (.756)	-.023 (.925)	-.001 (.996)	.173 (.466)	-.334 (.15)	-.003 (.989)
Recall	-.061 (.811)	-.19 (.45)	-.164 (.521)	**.514* (.05)**	**-.475* (.046)**	-.197 (.433)	-.249 (.335)

Table 6.25: Correlations between Groupings C and B

6.4.6 Discussion Points for Grouping C

Readability & Comprehension

The results for the human evaluation found no significant differences between the uncontrolled and controlled conditions in terms of their perceived readability and comprehensibility. There was, however, a significant difference between the uncontrolled and controlled conditions in terms of scores for items recalled, where the controlled condition resulted in significantly higher scores for all three variables. The correlational results of the human evaluation appear to find similar results as the more objective method of measuring readability via the traditional indices, but it should be remembered that textual measures such as Flesch and LIX do not explicitly measure comprehension and recall. This point relates back to the discussion in Chapters Two and Three with regard to the fuzzy definition in the literature concerning readability and its associates. Perhaps the operational definitions used in this study for readability and comprehensibility differed in such a way from the readability indices that the latter could not support the findings of the human evaluations of readability and comprehensibility.

Recall Testing

An additional issue is the development and testing of recall. As described in Chapters Two and Three, many different methods of testing recall have been employed and consistency and validity are of concern when adopting such methods. It would appear from the results in the previous sub-section showing a strong correlation between the scores of the comprehensibility evaluation and recall test scores, that the recall testing achieved its aim in this context.

Finally, an interesting parallel was found in that recall scores correlated with Flesch scores. Such a finding provides an interesting view of readability as, to the knowledge of the researcher, the interaction between Flesch and recall (as operationalised here) has not been studied before. Such a result is logical given that better Flesch scores aim to indicate better readability and, consequently,

recall. However, given the poorer correlation with the scores of the human evaluation of readability and comprehensibility, such a conclusion should not be so hastily drawn. Additionally, the correlation between readability and regression distance provides another interesting point and would appear logical in that for areas of the text that were deemed to be poor in readability as rated by the human evaluation, more regressions were made. However, scores for the human evaluation of comprehensibility did not support this argument.

6.4.7 Section Summary

This section examined the retrospective measures of readability, comprehensibility and recall, using data from a post-task human evaluation. Firstly, with regard to readability evaluation scores, no significant difference between the uncontrolled and controlled conditions was found. This was also the case for the comprehensibility evaluation, but not for recall scores, where the controlled condition was found to yield a significantly higher score. Secondly, a correlational analysis found significant strong correlations between readability evaluation scores and comprehensibility evaluation scores, as well as between comprehensibility and recall. Thirdly, when compared with the previous two groupings, a significant strong correlation was found between recall scores and Flesch from the textual variables of Grouping A and a significant moderate correlation was found between readability and regression distance from Grouping B. However, comprehensibility had no significant correlations with any variables from either Grouping A or B, while recall had significant moderate correlations with both PCPD and the number of regressions.

6.5 Regression Analysis

6.5.1 Section Overview

This section describes the multiple regression analysis used to identify possible predictors of readability and comprehensibility as measured by traditional readability indices, post-task human evaluation of readability and comprehensibility and recall testing. As described in Chapter Five, the analysis creates a model which incorporates one variable at a time, starting with the variable it has identified as having the most effect or predictive power, i.e. the predictor that accounts for most of the variance within the criterion variable. It should be noted that although the minimum requirements were met with the sample size in the study, a larger sample would result in a more accurate and valid model, and therefore, more generalisable results. As there is a limit to the number of potential predictor variables that can be used (approximately five times the number of cases than predictors are required) each grouping was entered separately and in accordance with the groupings A, B, and C above. This was also logical from a conceptual point of view in that the eye tracking variables were measured as a grouping, as were the textual variables and so on. This meant that for predicting human readability scores, the variables from the comprehensibility evaluation and the recall test were not used, e.g. it would not be of much value to know if a high score on a comprehensibility test predicts a high score of readability, especially as it has already been shown they correlated very strongly. In other words, this analysis attempts to find predictors of readability and comprehensibility from each of the other groupings of variables.

6.5.2 Textual Variables (Grouping A)

Firstly, to predict Flesch scores, the model, Model 1, which had an adjusted *r* square of .274 where $F = 5.145$ and $p = .047$, found observation length to be the best and only significant indicator with a β value of -.583, where $p = .047$. Figure 6.21 shows this result. Observation length could account for 27.4% of variance in Flesch scores, a weak but significant result.

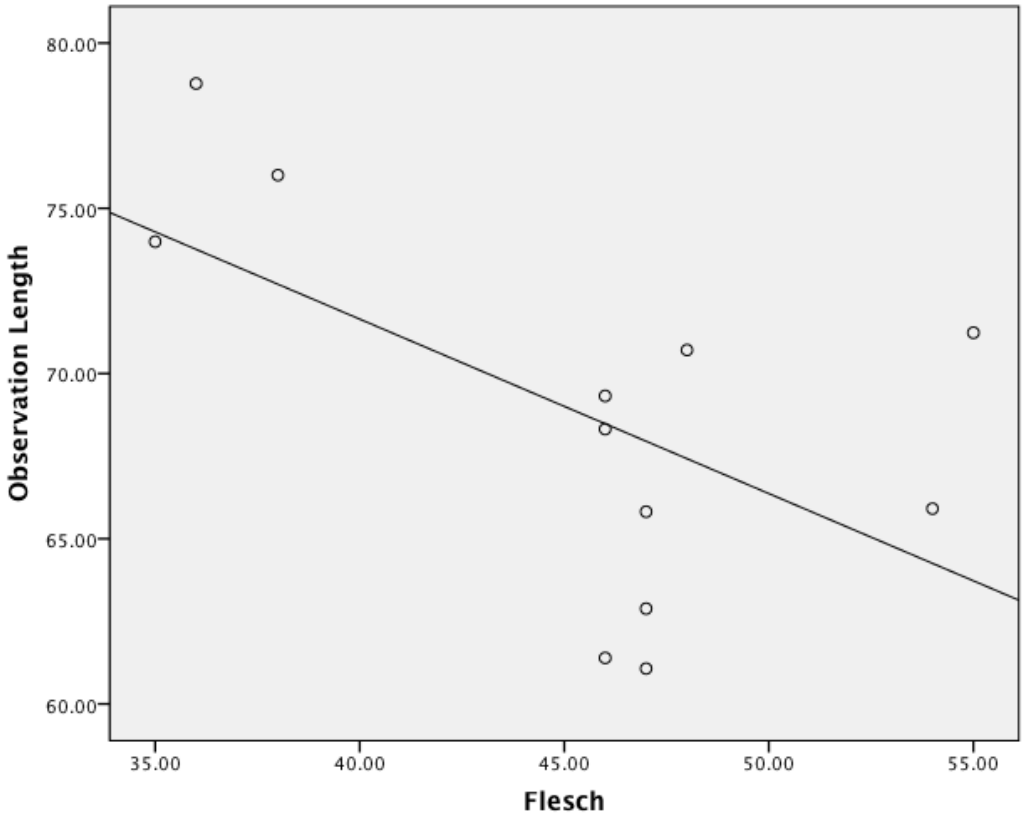

Figure 6.21: Linear Regression for Flesch and Observation Length

Secondly, in the prediction of LIX scores, the model, Model 1 with an adjusted *r* square of .362, $F = 7.242$, $p = .023$, found that percentage change in pupil dilation (PCPD) to be the best and only significant predictor with a β value of .648, $p = .023$. Percentage change in pupil dilation could account for 36.2% of variance in LIX scores, a moderately strong result. Figure 6.22 shows this result.

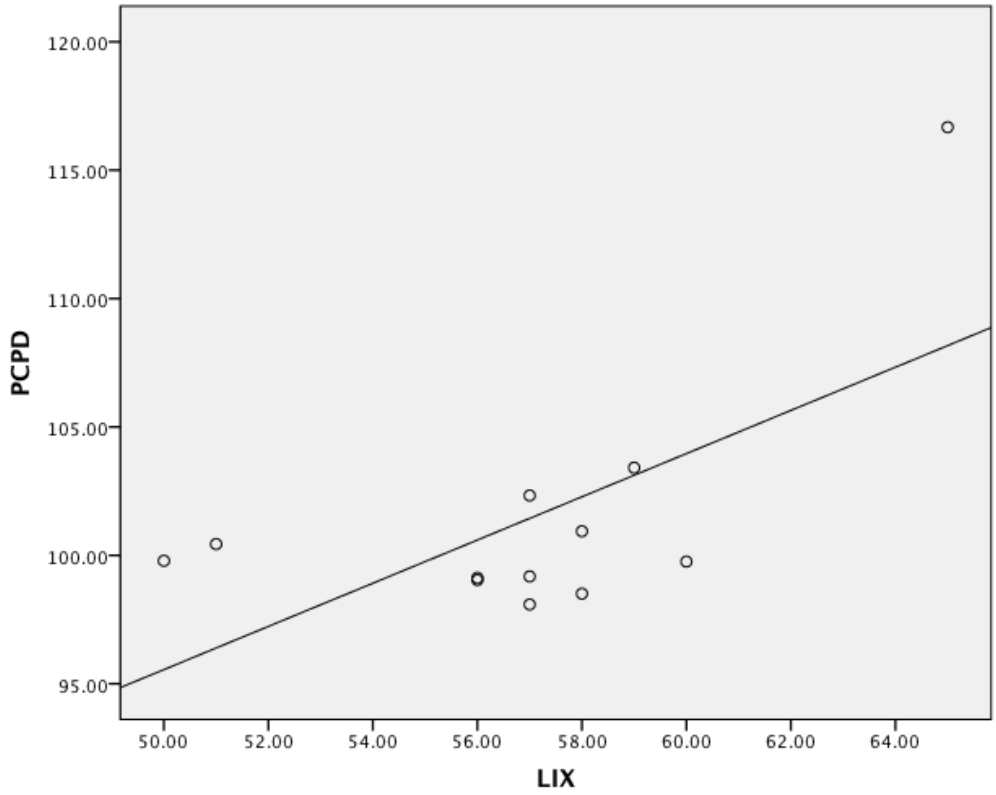

Figure 6.22: Linear Regression for LIX and PCPD

Finally, no significant predictors could be found for any of the AEMs: GTM, TER, or BLEU.

6.5.3 Human Evaluation Variables (Grouping C)

Firstly, for the prediction of the scores from the human evaluation of readability, the model, Model 2, had an adjusted *r* square of .398 with $F = 7.282$, and $p = .005$. Regression distance ≥ 10 was found to be the best predictor with a β value of -.838, $p = .001$. The number of regressions was found to be the next best predictor with a β value of .543, p = .024. Together both of these variables accounted for 39.8% of the variance within the readability variable; a moderate result. Figures 6.23 and 6.24 illustrate the regression lines for both predictors separately on the criterion of human readability evaluation scores. Secondly, for the prediction of the scores from the human evaluation of comprehensibility, no significant predictors could be found. This was also the case for recall scores.

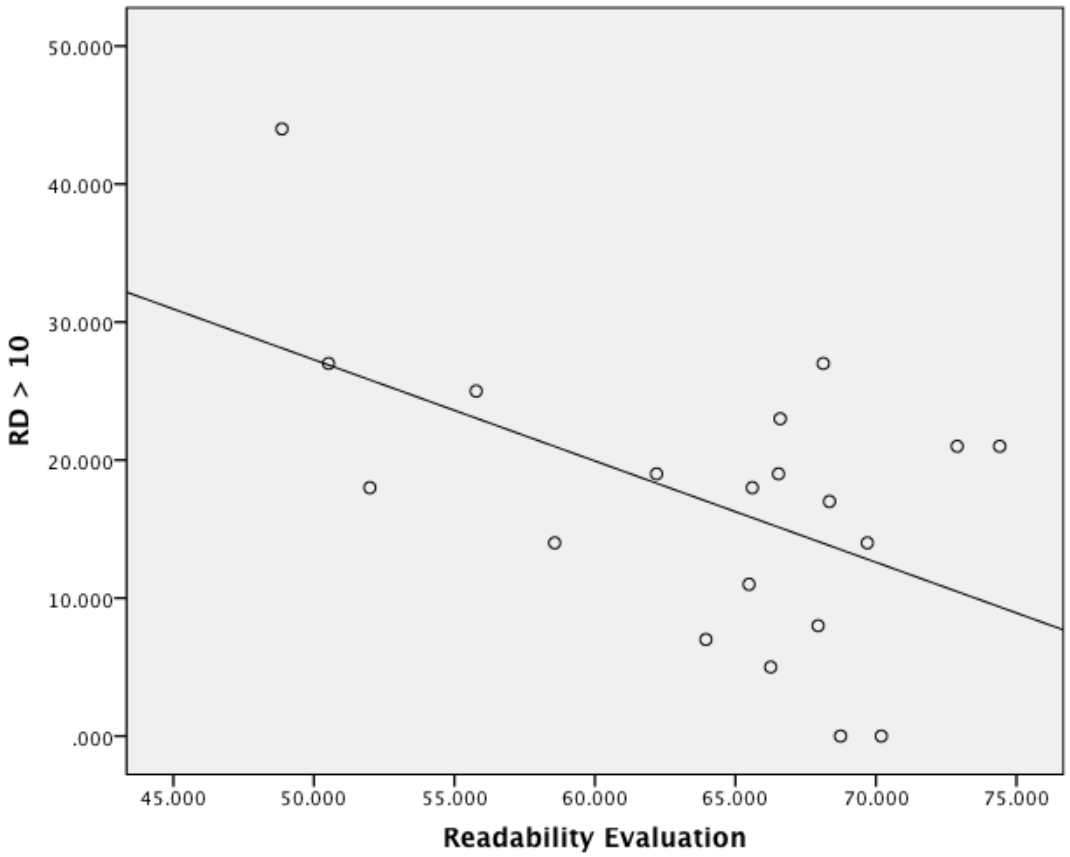

Figure 6.23: Linear Regression for Readability Evaluation and Regression Distance ≥ 10

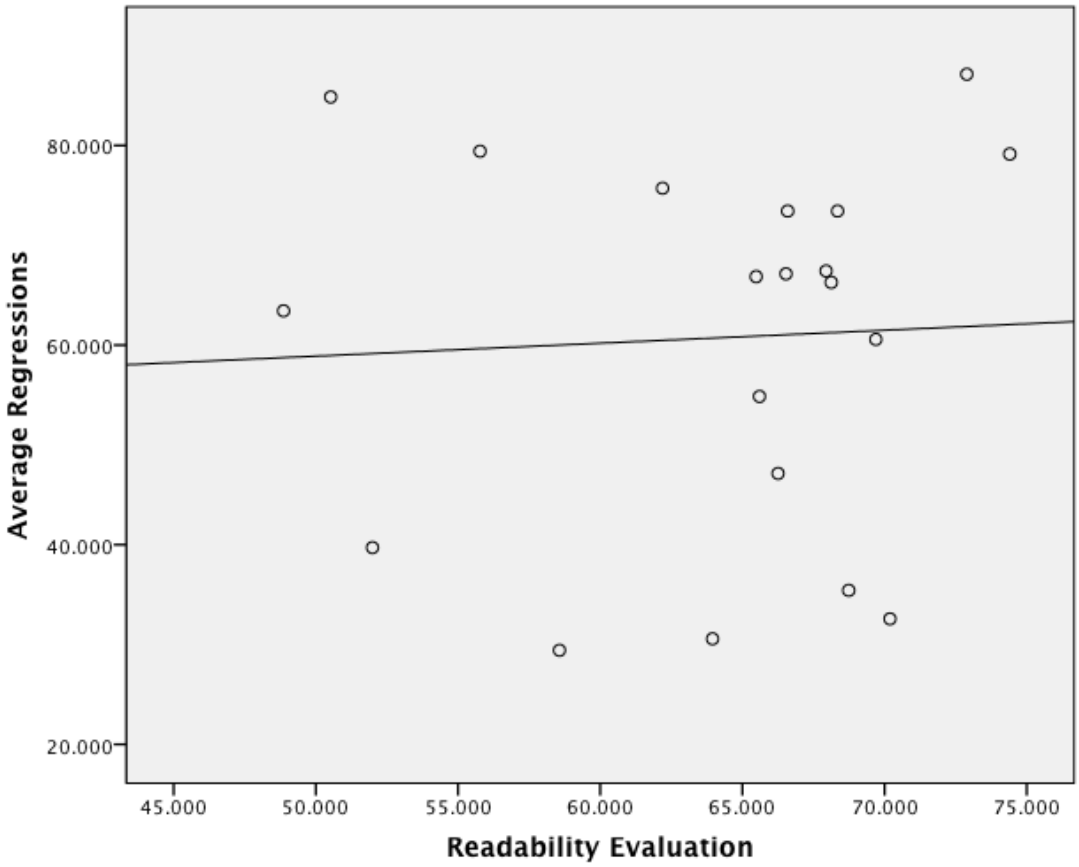

Figure 6.24: Linear Regression for Readability Evaluation and Number of Regressions

6.5.4 Section Summary

This section explored the use of stepwise multiple regression analysis to identify predictors of readability and comprehensibility as measured by the traditional readability indices of Flesch and LIX, as well as the scores from the human evaluations of readability, comprehensibility, and recall. For prediction of Flesch scores, observation length was found to be the best predictor, while for LIX scores, PCPD was identified. Moving to the human evaluation scores for readability, both regressions, and regression distance ≥ 10 were significant predictors. However, for the human evaluation scores for comprehension and the scores from the recall testing, no predictors could be identified.

The above results are said to be true in both directions so that Flesch can be said to be a predictor of observation length and LIX a predictor of PCPD. As long as observation length and PCPD are considered as indicators of cognitive effort, this means that easily computed Flesch and LIX readability scores can predict the cognitive effort involved in reading a text.

6.6 Validation of Hypotheses

To draw the findings presented in this chapter back to the research questions of the current study, this section lists the underlying null hypotheses of the research questions and serves as a summary of the implications of the above results. The main research question driving the study was:

- *Does the implementation of linguistic pre-processing in the form of a controlled language rule set result in higher levels of readability and comprehensibility in Statistical Machine Translation output?*

Embedded research questions asked:

- *Does implementation of CL result in improved scores as measured by the traditional readability indices Flesch and LIX?*
- *Are differences in eye tracking measures reported between the uncontrolled and controlled conditions?*
- *Do post-task human evaluation and recall testing show an improvement in readability and comprehensibility after implementation of CL?*
- *Do all of the above measures correlate and yield consistent findings?*
- *What is the relationship between human and machine evaluation of MT in this context?*

<u>Grouping A - Textual Measures</u>

The null hypotheses for metrics in Grouping A: Textual Measures were:

H1a: No significant difference would be found between conditions as measured by the Flesch and LIX readability indices.

The null was not rejected as there was no significant difference found; however, the controlled condition did result in slightly higher Flesch and LIX scores. The small sample size adds to the difficulty in finding a significant result.

> *H1b: No significant difference would be found between conditions using the automatic evaluations metrics of TER, GTM, and BLEU.*

The null was not rejected as there was no significant difference.

> *H1c: There would be no significant correlation between the readability indices and automatic metrics.*

The null was not rejected as there was no significant difference.

Grouping B - Eye Tracking Measures

The null hypotheses for metrics in Grouping B: Eye Tracking Measures were:

> *H2a: Average observation length would not be significantly different between the two conditions.*

The null was not rejected as no significant difference was found.

> *H2b: Average fixation count would not be significantly different between the two conditions.*

The null was rejected as a significant difference was found in that the controlled condition resulted in significantly fewer fixations.

> *H2c: Average fixation length would not be significantly different between the two conditions*

The null was rejected as a significant difference was found in that the controlled condition had significantly shorter average fixation lengths.

H2d: The average number of regressions would not differ significantly between conditions.

The null was not rejected as no significant difference was found between the conditions in terms of the number of regressions.

H2e: The average regression distance would not differ significantly between conditions.

The null was not rejected. However, when the measure used median values, a significant difference was observed.

H2f: Regression distance equal to or greater than ten would not differ significantly between conditions.

The null was rejected as a significant difference was found in that the controlled condition had significantly fewer regressions with a distance equal to or greater than ten words.

H2g: There would be no significant difference in pupil dilation between the two conditions.

The null was not rejected as no significant difference was found. Surprisingly the controlled condition resulted in a slightly higher average pupil diameter.

<u>Grouping C - Human Evaluation</u>

The null hypotheses for metrics in Grouping C: Human Evaluation were:

H3a: There would be no significant difference between conditions as measured by the human evaluation of readability.

The null was accepted as no significant difference was found between the conditions.

H3b: There would be no significant difference between conditions as measured by the human evaluation of comprehensibility.

The null was accepted as no significant difference was found between the conditions.

H3c: There would be no significant difference between conditions as measured by the retrospective recall testing.

The null was rejected as there was a significant difference in that the controlled condition resulted in a significantly greater score on the recall test.

Additional Hypotheses

H4a: No significant predictor of readability as determined by readability indices

The null was rejected as Flesch scores were found to be predicted by observation length, and PCPD was found to predict LIX scores.

H4b: No significant predictor of readability as determined by human evaluation

The null was rejected as both regression distance ≥ 10 and number of regressions were found to predict the human evaluation scores.

H4c: No significant predictor of comprehensibility could be identified.

The null was not rejected as no predictor could be found.

H4c: No significant predictor of comprehensibility could be identified.

The null was not rejected as no predictor could be found.

6.7 Chapter Summary

This chapter has presented the results of the main study and provided comparisons with other relevant studies in the literature. The results were presented in three groupings according to the nature of the variables in question: textual, eye tracking, and human evaluation.

From Grouping A, the textual variables, it was found that the uncontrolled output had more errors than controlled and a slightly different distribution of errors across sub-categories. Although Flesch and LIX scores were better for the controlled output, this increase was not statistically significant and a strong correlation between these metrics serves as support for their construct validity. No significant difference was found in terms of the AEMs (GTM, BLEU, and TER), and there was again a strong correlation between these measures. Yet, no significant correlations were found between Flesch/LIX and the AEMs highlighting the different constructs they evaluate. The discussion of textual variables also highlighted issues in sample size for textual measures and the problem of securing adequate reference translations to allow valid comparisons with MT output given different systems, i.e. controlled and uncontrolled.

For Grouping B, the eye tracking variables, no significant difference was found between conditions in terms of observation length, or PCPD. Fixation count and length were both found to be significantly lower in the controlled condition. With regard to the measure of regression, the number of regressions was not significantly different between conditions and nor was mean regression distance. A significant difference was found between conditions in terms of regression distance (using median scores) and the number of regressions ≥ 10, however.

It was also found that observation length correlated significantly with fixation count, which, in turn, also correlated significantly with fixation length. PCPD had no correlate within this grouping. Mean and total regressions were found not to correlate with each other. However, using the median value was found to result in a significant correlation with number of regressions, and regression distance ≥ 10 correlated with the mean regression distance, but not with either of the other two measures. When comparing the eye tracking

variables with those from Grouping A, it was found that observation length had a significant correlation with Flesch and PCPD, the latter was also a significant correlate of LIX.

The final grouping of variables, Grouping C, consisted of the post-task human evaluation and recall testing. A significantly higher score was found for the controlled condition for all three measures: human-evaluated readability and comprehensibility, and recall. Both readability and comprehensibility, and comprehensibility and recall were found to be significantly correlated. In addition, a significant strong correlation was found between recall score and Flesch from Grouping A of textual variables. From Grouping B, a significant moderate correlation was found between readability and regression distance. However, comprehensibility had no significant correlations with any variables from either grouping. Recall had significant moderate correlations with both PCPD and regressions, however.

The final section of the findings described the results of a multiple regression analysis which found observation length to predict Flesch scores, PCPD to predict LIX scores, and regression and regression distance ≥ 10 to predict human evaluation scores of readability.

Overall, it would appear that the more objective textual measures did not reveal the implementation of the CL to be of significance, and nor did several of the eye tracking variables. Other eye tracking variables did reveal significant differences between uncontrolled and controlled conditions, however. These variables were: fixation count, fixation length, and median regression distance. In complete contrast to the objective measures, the human evaluation deemed controlled MT output significantly more readable and comprehensible than uncontrolled output, and the human evaluators recalled significantly more items from the controlled output.

It would appear that in the eyes of the reader, figuratively and literally, adjustments are being made to compensate for problems in the text. Objective measures would, of course, highlight such errors, yet perhaps human readers are not as consistent as objective measures or maybe they are more willing to allow for errors, especially in the knowledge that the text is a translation in the first place (although participants were not aware that the text was a machine

translation). Moreover, such compensations may not be conscious and, therefore, may not be available to report via thinking aloud or retrospective evaluation tasks, while eye tracking metrics may reveal evidence of cognitive processes involved in such compensation and provide more insight into differences in participants' behaviour. This and other issues will be explored further in the conclusions.

Part IV:

Conclusion

Chapter Seven:

Conclusion

7.1 Research Aims

The main aim of this study was to answer the question:

- *Does the implementation of linguistic pre-processing in the form of a controlled language rule set result in higher levels of readability and comprehensibility in Statistical Machine Translation output?*

It has been demonstrated that certain measures (human evaluation and eye tracking) found significantly higher levels of readability and comprehensibility in the CL output, while other measures (Flesch and LIX indices) did not. It was argued that due to the small number of paragraphs on which the readability indices were employed, it was difficult to reach a significant result, even if an overall increase in readability for both the Flesch and LIX indices was observed. Nevertheless, it is apparent from the results of the study that human readers are more sensitive to the effects of CL, and consequently so are measures that directly or indirectly measure the human reading/evaluation process.

On the other hand, objective measures such as Flesch and LIX indices are not as sensitive to CL because they count a limited number of text attributes (e.g. number of words, number of syllables) and remain indifferent to other attributes to which humans are more sensitive. The overall increase in Flesch and LIX scores observed in the study reported on here, however, can be attributed to the implementation of CL rules that have effects that are captured by these readability indices. For example, the rule that affects sentence length has an effect on both indices while other rules, such as the one that proscribes the use of passive voice (e.g. change "if your installation is managed by your administration" to "if your administrator manages your installation") may have a minimal impact on the indices but a greater effect on the human reader.

Additional related questions were derived from the main research question:

- *Does implementation of CL result in improved scores as measured by the traditional readability indices Flesch and LIX?*

It was demonstrated that the readability indices of Flesch and LIX did not show a significant improvement in scores for the text edited in accordance with the CL rule set. As stated, due to the small number of paragraphs used in the study (six), a significant finding was unlikely. Nevertheless, a slight overall improvement was identified in the controlled text. It may be the case that the rules unique to the rule set used in this study do not engage with the items measured by the indices as such, e.g. number of syllables, and further testing is advised (see below).

- *Are differences in eye tracking measures reported between the uncontrolled and controlled conditions?*

The findings from the eye tracking data indicated that several of the measures resulted in significantly *better* scores which are interpreted as indicating lower cognitive effort involved in reading, i.e. the controlled condition had fewer fixations, shorter fixation lengths, shorter median regression distance, and fewer regressions ≥ 10 words. However, four (of the eight) eye tracking measures did not find significant differences between the conditions: observation length, percentage change in pupil dilation, and mean number of regressions. Given the lack of convergence between different eye tracking measures noted in the previous chapter, and given the wide variety of eye tracking metrics used in this study relative to comparable studies, it is perhaps not surprising to find a lack of complete convergence in the results presented here for all eight eye tracking metrics.

- *Do post-task human evaluation and recall testing show an improvement in readability and comprehensibility after implementation of CL?*

The human evaluation did not result in significant differences between conditions for perceived readability and comprehensibility, even though the controlled condition had higher scores for both. However, a significant difference was observed for recall scores, where the controlled condition resulted in significantly higher scores. In addition, both conditions rated sentences as more comprehensible than readable, which may indicate the readers' ability to compensate for linguistic errors in the text, e.g. those described in the error analysis in Chapter Six. Such errors may be penalised strongly by other forms of evaluation such as readability indices and automatic evaluation metrics and this may serve to explain the difference found between the subjective and (more) objective measures employed in the study.

- *Do all of the above measures correlate and yield consistent findings?*

Many of the measures correlated well with similar measures but not to the same extent with measures of a different nature. For example, the readability indices of Flesch and LIX correlated with each other, as did each of the automatic evaluation metrics, yet neither Flesch nor LIX correlated with any of the latter. In terms of eye tracking, several correlations were found, namely: fixation count with fixation length and observation length, median regression distance with number of regressions and regressions ≥ 10 words. However, pupil dilation did not correlate with any other measures within the same grouping. It can be concluded that pupil dilation is not a valid measure of cognitive effort as operationalised in the current study. Finally, as stated above, a very strong correlation was found between the human evaluation of readability and comprehensibility, and between comprehensibility and recall.

- *What is the relationship between human and machine evaluation of MT in this context?*

It was found that human evaluators favoured the controlled MT output: they judged it more readable and comprehensible, and they also retained significantly more of the content of the controlled output, as measured by the recall test. In contrast, automatic evaluation metrics favoured the uncontrolled text, but not significantly so. The need to use an appropriate human reference translation was presented as a likely cause for this result, especially as the automatic evaluation metrics used in the study have been reported elsewhere to correlate well with human evaluation. In this study, however, they did not correlate to a significant extent with human judgements of readability and comprehension.

7.2 Practical Implications for the Implementation of CL in MT Workflows

Bringing together the above results and the overall findings of the study, it can be concluded that the implementation of CL is likely to result in quantifiable improvements in MT output in terms of readability and comprehensibility, especially where human end-users, as opposed to natural language processing applications, are concerned. In the context of CL and MT, several indicators of readability were identified using a resource-cheap means of measuring readability, i.e. the Flesch index predicted observation length (so better Flesch scores would result in shorter observation lengths), and LIX predicted changes in pupil dilation, thereby suggesting it is more closely related to cognitive processing than Flesch. However, as the measure of pupil dilation did not correlate well in other analyses, this conclusion is tentative and would require further study. Further to this, it was also found that recall correlated significantly with pupil dilation, regressions, and Flesch scores, while the human evaluation of readability correlated with regression distance.

Overall, it would appear that the implementation of CL offers strong potential when used in conjunction with MT. However, such an implementation would have to be carefully considered taking into account the needs of the organisation and factors such as readership, language pairs, MT systems, and methods of evaluation. Given the cost of human evaluation, it may be advisable to conduct evaluation at intermediate stages via readability indices, automatic evaluation metrics, and small-scale eye tracking and usability studies, before a large-scale human evaluation would take place, for example, when the system/rule set has demonstrated sufficient improvements via the above measures.

7.3 Limitations

Sample Size

Although a strength of the study was its relatively large sample size when compared to other eye tracking studies in translation process literature, a final number of twenty participants is still relatively small for robust statistical analyses. Therefore, the sample size is noted as a limitation of the study and can be attributed to:

i. the large amounts of data produced using the eye tracking method, much of which required wholly manual analysis;
ii. the need to involve participants who were native speakers of the target language (French) and who had little to no prior domain knowledge;
iii. the ethical requirements for human research participants to be volunteers;
iv. the need for the eye tracker to remain on-site at the research laboratory both for the sake of environmental continuity and due to its relative immobility (O'Brien 2008).

Materials

The question of ecological validity was of concern when planning the main and pilot studies especially with regard to the materials used, namely the corpora, the CL rule set, and the MT system. Fundamentally, ecological validity is concerned with whether the research question requires use of materials that are genuinely used outside the confines of a research environment. Where artificial materials are used, e.g. a corpus seeded with a number of instances of a CL rule violations to test its effects, there is scope for more fine-tuned and possibly more precise findings. On the other hand, using 'naturally occurring' corpora and CL rules, i.e. materials that are currently in use in a commercial environment, provides valuable insights into real-life scenarios surrounding the research context. A limitation is therefore inevitable when adopting one approach over

the other, and it would have been valuable to highlight specific CL rules and investigate their effect in isolation. However, this was beyond the scope of the current study but is highlighted as a point for further research (see section 7.5 below).

As the study was designed to be reflective of real reading of the text, it was not appropriate to control participants' eye movements as seen in various eye tracking experiment methodologies (e.g. McConkie and Rayner 1975, Jensen 2009). Such control may have resulted in more accurate data or, at the very least, the loss of fewer participants due to poor data quality.

Regarding the use of readability indices, a limitation existed as the paragraphs in the current study were in French. As described in Chapter Two, the vast majority of research into readability indices applies only to English. Novel approaches that incorporate deeper linguistic information, for example, the Coh-Metrix measure (Graesser *et al*. 2004), are potentially more promising but at the time of the study such measures were not available.

The language pair itself is another limitation in that the findings may be influenced by the nature of the properties of and relationship between English and French. However, given the scope of measures used in the study, it was not possible to conduct several experiments to cover other languages. Furthermore, the choice of language pair was also limited by the corpora and thus MT system available and the language expertise of the researcher and the available participants.

Lastly, it would be valuable to replicate the study using other MT systems, and indeed other paradigms of MT, for example, RBMT (see section 7.5 below).

Experiment Design

The design of the experiment presents several limitations in that a randomised design, as adopted in the pilot study, may have been of value to the main study. On the other hand, the presentation of coherent full texts was appropriate in the main study as paragraphs had to be at least one hundred words in length due to the use of readability indices, whereas single out-of-context sentences were used in the pilot.

As described in Chapter Two, while the study acknowledges additional variables such as: reader type, level of language competence in the native (in this case French) and foreign language (English), and differences in levels of working memory, it did not account for these variables, although they have been shown to have an effect on the cognitive processes and phenomena that take place during reading, e.g. motivation (Shnayer 1969, Carrell 1987, Schriver 1989).

The complexity of data generated during eye tracking and the time required for their analysis presents another limitation. While some studies suggest using more automated means of analysing data, such as using a scripting language like Perl (Jensen 2009), much of the analysis requires more intuitive judgement. Although automated means are, of course, extremely useful in terms of the amount of time and effort saved, (subjective) human analysis and interpretation is still paramount as explained below.

A case in point is the counting of regressions and the distance regressed. While an algorithm could count the number of regressions, following fixations on other sentences and parts of the screen requires more complex interpretation as sometimes the reader's behaviour is ambiguous and it may take several replays to ascertain, albeit subjectively, what the reader was doing. Having said that, researchers such as Carl (2008) have begun to explore the potential for more automation in eye tracking studies. Likewise, more sophisticated mixed and regression methods have been demonstrated in studies comparable to this one (e.g. Balling 2008). However, manual inspection of eye tracking data to ensure accuracy is still advisable at several stages throughout an eye tracking study (Jensen 2009).

7.4 Contributions

Despite the aforementioned limitations, it is believed that the study adequately answered the proposed research questions and added to the body of knowledge that exists on these topics.

As noted by Jensen (2009), few studies have used larger text units such as whole texts or even paragraphs for studies of cognitive processing. While the pilot study examined isolated sentences from the same corpus, the main study moved up to full and in-context paragraphs and, in this sense, it goes some way towards addressing the scarcity of studies that examine larger units of flowing and coherent texts.

As mentioned in the previous sub-section, the sample size was a limitation; however, and as highlighted elsewhere (O'Brien 2010, Hvelplund 2011), it was well above the typical size used in eye tracking studies and this provides support for the generalisablity and validity of the findings.

The ecological validity of the study and consequently of its results is believed to be a strength in that the findings are based on corpora that are actually used in MT workflows. This makes any recommendations for the implementation of CL in conjunction with MT more compelling from the point of view of MT developers, industrial users, and other interested parties.

The novel approach of the study combined eye tracking with other measures of readability and comprehensibility, a method which has hitherto been unexplored in the area. Like other eye tracking studies, it highlights the value of mixed-method designs and the complementary nature of appropriate research methods. The mixed-method design combined both quantitative and qualitative approaches to address the research questions and proved especially useful in the investigation of, for example, the differences between human and automated evaluation of MT, a topic that has been a constant issue in the literature (Papineni *et al*. 2002, Banerjee and Lavie 2005, Snover *et al*. 2006). In addition to this, to the knowledge of the researcher, there are no other studies that investigate the use of CL in the context of SMT.

Furthermore, it employed several indicators of readability, comprehensibility, and cognitive effort, which provided a more robust and

comprehensive approach to answering the research questions, while ascertaining the value, validity, and limitations of each indicator both on its own and in conjunction with others. Specifically, the research reported on here identifies a number of correlations between phenomena whose interactions have not been investigated before. These include correlations, for example, between Flesch scores and human recall, and human judgement of readability and regressions made during reading.

7.5 Future Research

Replication of the study or its components would be an important task for future work. Specifically, it would be valuable to ascertain the effect that each CL rule has on the readability and comprehensibility of a text, thus forming a more comprehensive set of guidelines for users of CL, especially as needs of users vary to a large extent. In addition, such a study should also be extended in terms of the language pairs used, the text types, and MT systems.

Combination of eye tracking with electroencephalography (EEG) or functional magnetic resonance imaging (fMRI) methods of concurrent process analysis would provide extremely rich results. Given their shortcomings, EEG and fMRI methods would benefit from the integration of eye tracking; for example, typical EEG set-ups identify activation of the optic nerve and streaming of visual information through it, but they cannot identify eye behaviour or areas of focus etc. on screen. By combining methods, researchers could get more robust measures of cognitive effort and possibly reconcile confounding results in eye tracking measures, such as pupil dilation, in relation to cognitive processing in a given context (cf. Chang 2009).

Additionally, while controlling the input to the MT process has been consistently demonstrated to yield better output, other similar methods have been explored, such as controlling the output through controlled generation (Way and Gough 2003) by means of restricting the population of the phrase table in the MT process vis-à-vis a set of predefined rules. Such methods also present fruitful avenues of further research.

Furthermore, it is believed that the current study was successful in bringing together approaches from several domains that could be of mutual benefit. While the establishment of CL guidelines is an ideal milestone, many variables need to be addressed given different user needs, resources etc. However, features of CL could be classified and used in many processes such as controlled authoring of content and its translation both by MT and using translations memories and other computer-aided translation tools.

Additionally, it should be noted that CLs are currently often applied just in the writing of source-language content; but later users of this content, for

example, translators or post-editors who by definition write in the target language could, in future, also be required to adhere to CL guidelines. Currently the advantages of CL implementation (e.g. reduced ambiguity, and improved readability and comprehensibility of the content) may be lost or diminished as texts move through complex workflows. Further work on maintaining CL-inspired consistency in these scenarios would be welcome.

Finally, building upon the measurement of readability and comprehensibility of MT output, a move to the study of usability, especially in the context of technical support documentation, represents a potentially effective means of measuring the user experience of machine-translated text. It would also enable researchers to address issues such as user/reader motivation, especially when real users of the content are included in the evaluation process.

Reference List

Adelberg, A.H. and Razek, J.R. 1984. The Cloze procedure: a methodology for determining the understandability of accounting textbooks. *The Accounting Review*, 59 (1), pp. 109–122.

Adriaens, G. 1994. Simplified English Grammar and Style Correction in an MT Framework: the LRE SECC Project. IN: *Proceedings of Translating and the Computer* 16, London, pp. 78-88.

Agarwal, A. and Lavie, A. 2008. Meteor, M-BLEU and M-TER: evaluation metrics for high-correlation with human rankings of machine translation output. IN: *Proceedings of the Third Workshop on Statistical Machine Translation, 19 June, Columbus, Ohio*, pp.115-118.

Aikawa, T., Melero, M., Schwartz, L. and Wu, A. 2001. Generation for multilingual MT. MT Summit VIII: *Proceedings of Machine Translation in the Information Age, Santiago de Compostela, Spain, 18-22 September*, pp. 9-14.

Aikawa, T., Schwartz, L., King, R., Mo, C.O. and Lozano, C. 2007. Impact of Controlled Language on Translation Quality and Post-editing in a Statistical Machine Translation Environment. IN: *Proceedings of MT Summit XI, Copenhagen, Denmark*, pp. 1-7.

Allen, J. 1999 Adapting the concept of "Translation Memory" to "Authoring memory" for a Controlled Language writing environment. IN: *The Proceedings of the 21st Conference of Translating and the Computer 21, 10-11 November 1999.* London: ASLIB, pp. 192-199.

Allen, J. 2003. Post-editing. IN: H. Somers (ed.). *Computers and Translation: A Translator's Guide.* Amsterdam: John Benjamins Publishing Company, pp. 297-1317.

Almqvist, I. and Hein, A. 1996. Defining ScaniaSwedish - a Controlled Language for Truck Maintenance. IN: *Proceedings of the First International Workshop on Controlled Language Applications (CLAW 96), Leuven, Belgium*, pp. 159–164.

Altarriba, J., Kroll, J. E, Sholl, A., and Rayner, K. 1996. The influence of lexical and conceptual constraints on reading mixed language sentences: Evidence from eye fixation and naming times. *Memory and Cognition*, 24, pp. 477-492.

Alves, F. (ed.). 2003. *Triangulation Translation: Perspective in Process Oriented Research*. Amsterdam and Philadelphia: John Benjamins.

Alves, F., Pagano, A., Silva, I. A. 2009. New Window on Translators' Cognitive Activity: Methodological Issues in the Combined Use of Eye Tracking, Key Logging and Retrospective Protocols. IN: Mees, I., Alves, F., Göpferich, S. (eds.). *Methodology, Technology and Innovation in Translation Process Research: A Tribute to Arnt Lykke Jakobsen. Copenhagen Studies in Language 39*. Samfundslitteratur: Copenhagen.

Anastasiou, D. 2008. Identification of idioms by machine translation: a hybrid research system vs. three commercial systems. IN: *Proceedings of the 12th Annual Conference of the European Association for Machine Translation, September 22 and 23, 2008, Hamburg, Germany*, pp. 12-20.

Anderson, J. 2000. *Cognitive Psychology and its Implications* (5th ed.). New York: Worth.

Andersson, B., Dahl, J., Holmqvist, K., Holsanova, J., Johansson, V., Karlsson., Stromqvist, S., Tufvesson, S., and Wengelin, A.. 2006. Combining Keystroke Logging with Eye-tracking. IN: Van Waes, L., Leijten, M., and Neuwirth, C. (eds.), *Writing and Digital Media*, 17, pp. 166-172.

Andonova, E., Janyan, A., and Popivanov, I. 2009. Brain activity and eye movements in translation. IN: *The Proceedings of the European Future Technologies Conference 21-23 April, 2009, Prague, Czech Republic,* no page numbers.

Aranberri Montasterio, N. 2009. *-ing Words in RBMT: Multilingual Evaluation and Exploration of Pre- and Post-processing Solutions*. PhD thesis. Dublin City University.

Armstrong, S., Way, A., Caffrey, C., Flanagan, M., Kenny, D. and O'Hagan, M. 2006. Improving the quality of automated DVD subtitles via example-based machine translation. Translating and the Computer 28. IN: *The Proceedings of the 28th International Conference on Translating and the Computer 28, 16-17 November 2006,* London: ASLIB, no page numbers.

Arnold, D. 2003. Why translation is difficult for computers. IN: Somers, H. (ed.). *Computers and Translation: A translator's guide,* pp. 119-142.

Atkinson, R. and Shiffrin, R. 1968. Human memory: a proposed system and its control processes. IN: Spence, K., and Spence, J. (eds.). *The Psychology of Learning and Motivation: Advances in Research and Theory, 2*. New York: Academic Press

Aue, A. Menezes, A, Moore, B. Quirk, C. and Ringger, E. 2004. Statistical machine translation using labelled semantic dependency graphs. IN: *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation, October 4-6, 2004, Baltimore, Maryland, USA*, pp. 125-134.

Avramidis, E. and Koehn, P. 2008. Enriching morphologically poor languages for statistical machine translation. IN: *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 15-20, 2008, The Ohio State University, Columbus, Ohio, USA*, pp. 763-770.

Baddeley, A. D. 2000. The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences 4* (11), pp. 417-423.

Baddeley, A. D. 2003. Working memory: looking back and looking forward. *Nature Reviews: Neuroscience, 4*, pp. 829-839.

Baddeley, A. D. and Warrington, E. K. 1970. Amnesia and the distinction between long- and short-term memory. *Journal of Verbal Learning and Verbal Behavior, 9*, pp. 176-189.

Baddeley, A. D. and Hitch, G. J. 1974. Working memory. IN: Bower, G. A. (ed.). *The Psychology of Learning and Motivation: Advances in Research and Theory, 8.* New York: Academic Press, pp. 47-89

Balling, L. W. 2008. A brief introduction to regression designs and mixed-effects modelling by a recent convert. IN: Göpferich, S., Jakobsen, A. Mees, I. (eds.) *Looking at eyes – eye tracking studies of reading and translation processing. Copenhagen Studies in Language 36.* Samfundslitteratur, Copenhagen, pp. 175-192.

Banchs, R. and Li, H. 2008. Exploring Spanish-morphology effects on Chinese-Spanish SMT. IN: *Proceedings of MATMT 2008: Mixing Approaches to Machine Translation, Donostia-San Sebastian (Spain), February 14th*, pp. 49-53.

Banerjee, S. and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. IN: *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan*, pp. 65-72.

Bartels, M. and Marshall, S. P. 2006. Eye tracking insights into cognitive modelling. IN: *Proceedings of the 2006 Symposium on Eye Tracking Research and Applications, San Diego, California, March 27 - 29, 2006, ETRA '06. ACM, New York.*

Barthe, K. 1998. GIFAS Rationalised French: Designing one Controlled Language to Match Another. IN: *Proceedings of the Second International Workshop on Controlled Language Applications, Pittsburgh, PA.*

Barthe, K., Juaneda, C., Leseigneur, D., Loquet, J. C., Morin, C., Escande, C. and Vayrette, A. 1999. GIFAS Rationalized French: A Controlled Language for Aerospace Documentation in French. *Technical Communication, 46* (2), pp. 220-229.

Beatty, J. and Lucero-Wagoner, L. 2000. The pupillary system. IN: Cacioppo, J. T., Tassinary, L. G., and Berntson, G. G. (eds). *Handbook of Psychophysiology* (2nd ed.) Cambridge: Cambridge University Press, pp. 142-163.

Beatty, J. 1982. Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin, 91*(2), pp. 276-92.

Bernth, A. 1997. EasyEnglish: A tool for improving document quality. IN: *Proceedings of the ACL 5th Conference on Applied Natural Language Processing, Washington, DC.*

Bernth, A. 1998. EasyEnglish: Pre-processing for MT. IN: Mitamura *et al.* (eds.), pp. 30-41.

Bernth, A. 1999. Controlling input and output of MT for greater user acceptance. IN: *The Proceedings of the 21st Conference of Translating and the Computer sponsored by ASLIB, 10-11 November 1999.* London: ASLIB, no page numbers.

Bernth, A. and Gdaniec, C. 2001. MTranslatability. *Machine Translation, 16*, pp. 175–218.

Bernth., A and McCord, M. 2000. The Effect of Source Analysis on Translation Confidence. IN: *Envisioning Machine Translation in the Information Future, 4th Conference of the Association for Machine Translation in the Americas, Springer Verlag, Berlin/Heidelberg*, pp. 89–99.

Betts, R. G. 2003. EasyEnglish: challenges in cross-cultural communication Controlled language translation. IN: *Proceedings of EAMT-CLAW-2003, Dublin City University, 15-17 May 2003,* pp. 8-15.

Bjork, R. A., and Bjork, E. L. 1992. A new theory of disuse and an old theory of stimulus fluctuation. IN: Healy, A., Kosslyn, S., and Shiffrin, R. (eds.). *From learning processes to cognitive processes: Essays in honor of William K. Estes, 2. Hillsdale, NJ: Erlbaum,* pp. 35-67.

Björnsson, C. H. 1968. *Läsbarhet*. Stockholm, Sweden: Bokförlaget Liber.

Björnsson, C.H. 1983. Readability of newspapers in 11 languages. *Reading Research Quarterly, 18* (4) pp. 476-484.

Blunsom, P., Cohn, T. and Osborne, M. 2008. A discriminative latent variable model for statistical machine translation. IN: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 15-20, 2008, The Ohio State University, Columbus, Ohio, USA*, pp. 200-208.

Bormuth J. R. 1966. Readability: A New Approach. *Reading Research Quarterly, 1*, pp. 79-132.

Bormuth, J. R. 1969. Experimental determination of the instructional reading level. IN: Figurel, J. A. (ed.). *Reading and Realism*. Newark Delaware: International Reading Association.

Bowker, L. and Pearson, J. 2002. *Working with Specialized Language: A practical guide to using corpora*. London/New York: Routledge.

Bowker, L. and Ehgoetz, M. 2007. Exploring User Acceptance of Machine Translation Output: A Recipient Evaluation. IN: Kenny, D., and Ryou, K. (eds.).

Across Boundaries: International Perspectives on Translation Studies. Newcastle, UK: Cambridge Scholars Publishing. pp. 209-224.

Bransford, J. D., and Johnson, M. K. 1972. Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, pp. 717-726.

Broadbent, D. E., Fitzgerald, P., and Broadbent, M. H. P. 1986. Implicit and explicit knowledge in the control of complex systems. *British Journal of Psychology, 77*, pp. 33-50.

Cadwell, P. 2008. *Readability and Controlled Language*. MA dissertation. Dublin City University.

Caffrey, C. 2009. Using pupillometric, fixation-based and subjective measures to measure the processing effort experienced when viewing subtitled TV anime with pop-up gloss. IN: Göpferich, S., Jakobsen, A. Mees, I. (eds.) *Looking at eyes – eye tracking studies of reading and translation processing. Copenhagen Studies in Language 36.* Samfundslitteratur, Copenhagen, pp. 125-144.

Cahill, A. 2009. Correlating human and automatic evaluation of a German surface realiser. IN: *Proceedings of the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing of the 47th Annual Meeting of the Association for Computational Linguistics, 2-7 August, Singapore*, pp. 97-100.

Callicott, J., Mattay, V., Bertolino, A., Finn, K., Coppola, R., Frank, J., Goldberg, T. and Weinberger, W. 1999. Physiological characteristics of capacity constraints in working memory as revealed by functional MRI. *Cerebral Cortex 9*, pp. 20-26.

Callison-Burch, C., Osborne, M. and Koehn, P. 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 EACL 2006, Trento, Italy, pp. 249-256.

Carl, M. 2003. Data-Assisted Controlled Translation. IN: *EAMT-CLAW 03, Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Controlled Translation, Proceedings, Dublin, Ireland.*

Carl, M. and Way, A. (eds.). 2003. Recent advances in example-based machine translation. *Text, Speech, and Language, 21.* Dordrecht/Boston/London: Kluwer Academic Publishers.

Carl, M., Jakobsen, A. and Jensen, K. 2008. Modelling human translator behaviour with user-activity data. IN: *Proceedings of 12th annual conference of the European Association for Machine Translation, September 22 and 23, 2008, Hamburg, Germany*, pp. 21-26.

Carrell, P. L. 1987. Readability in ESL. *Reading in a Foreign Language 4* (1), pp. 21-40.

Chall, J. S. 1958. Readability: An appraisal of research and application. *Bureau of Educational Research Monographs, 34*. Columbus: Ohio State University Press.

Chall, J. S., and Dale, E. 1995. *Readability revisited: The new Dale-Chall readability formula*. Cambridge, MA: Brookline Books.

Chan, Y., Ng, H. and Chang, D. 2007. Word sense disambiguation improves statistical machine translation. IN: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, June 2007*, pp. 33-40.

Charrow, R. and Charrow, V. 1979. Comprehension of Standard Jury Instructions: A Psycholinguistic Approach. *Columbia Law Review,* 79.

Chen, B., Zhang, M., Aw, A., and Haizhou, L. 2008. Regenerating hypotheses for statistical machine translation. IN: Proceedings: *Coling 2008: 22nd International Conference on Computational Linguistics, 18-22 August 2008,* Manchester UK, pp.105-112.

Chiang, D. 2005. A hierarchical phrase-based model for statistical machine translation. *ACL-2005: 43rd Annual meeting of the Association for Computational Linguistics, University of Michigan, Ann Arbor, 25-30 June 2005*; pp. 263-270.

Clark, J. H., Frederking, R. and Levin, L. 2008. Inductive detection of language features via clustering minimal pairs: towards feature-rich grammars in machine translation. IN: *Proceedings Second ACL Workshop on Syntax and Structure in Statistical Translation, 20 June 2008, Columbus, Ohio, USA*, pp. 78-86.

Clémencin. G. 1996. Integration of a CL-Checker in an Operational SGML Authoring Environment. IN: *Proceedings of The First International Workshop On Controlled Language Applications, Katholieke Universiteit Leuven, Belgium,* pp. 32-40.

Collins-Thompson, K. and Callan, J. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology, 56,* (13), pp. 1148-1462.

Connatser, B. R. 1999. Last rites for readability formulas in technical communication. *Journal of Technical Writing and Communication, 29* (3), pp. 271-287.

Converse, J. M. and Presser, S. 1986. Survey questions: Handcrafting the standardized questionnaire. *Quantitative applications in the social sciences.* Thousand Oaks, CA, US: Sage Publications.

Conway, M., Gathercole, S., and Cornoldi, C. 1998. *Theories of memory: Volume 2*. Psychology Press: East Sussex, UK.

Coolican, H. 1996. *Research Methods and Statistics in Psychology* (2nd ed.). London, England: Hodder and Stoughton Educational.

Coughlin, D. 2003. Correlating automated and human assessments of machine translation quality. IN: *Proceedings of MT Summit IX, 23-27 September, New Orleans, Louisiana*, pp.63-70.

Craik, F. I. M., and Lockhart, R. S. 1972. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, pp. 671-684.

Craik, F. I. M., and Tulving, E. 1975. Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, pp. 268-294

Craik, F. I. M., and Watkins, M. J. 1973. The role of rehearsal in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 12,* pp. 599-607.

Creswell, J. and Plano Clark, V. 2007. *Designing and conducting mixed methods research*. Thousand Oaks, California: Sage Publications.

Creswell, J. W. 2003. *Research Design: Quantitative and Qualitative Approaches* (2nd ed). Thousand Oaks, California: Sage Publications.

Crotty, M. 1998. *The foundations of social research: Meaning and perspective in the research process*. Thousand Oaks, California: Sage Publications.

Dale, E. and Chall, J. S. 1948. A formula for predicting readability. *Educational Research Bulletin, 27*, pp. 37-54.

Daneman, M., and Carpenter, P. A. 1980. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* pp. 450-466.

Danks, H. J. and Griffin, J. 1997. Reading and translation: a psycholinguistic perspective. IN: Danks, H. J., Shreve, G. M., Fountain, S. B., McBeath, M. K. (eds). *Cognitive Processing in Translation and Interpreting*. Thousand Oaks: CA: Sage, pp. 161-175.

Davison, A. and Kantor, R.N. 1982. On the failure of readability formulas to define readable text: A case study from adaptations. *Reading Research Quarterly, 17* (2), pp. 187-209.

De Bot, K. 2000. A Bilingual Production Model: Levelt's Speaking Model Adapted. *The Bilingualism Reader,* Psychology Press.

De Beaugrande, R., and W. U. Dressler . 1981. *Introduction to text linguistics*. London: Longman.

De Landsheere, G. 1963. Pour une application des tests de lisibilité de Flesch à la langue francaise, *Le Travail Humain, 26* (1,2), pp. 141-154.

De Landsheere, G. 1966 Lecteurs et lectures, re- cherches sur l'évaluation et le contrôle objectif. IN: *XIIe colloque international de I'Association de Pédagogie Experimentale de Langue Frangaise. Genève, 2-5: avril 1966*, pp. 139-165.

De Landsheere, G. 1973: *Le test de closure, mesure de la lisibilité et de la compréhension*. Bruxelles: Nathan-Labor.

De Preux, N. 2005. How much does controlled language improve machine translation results? IN: *The Proceedings of the 27th Conference of Translating and the Computer 27,* London: ASLIB, pp. 1-14.

Déchelotte, D.,Schwenk, H., Bonneau-Maynard, H., Allauzen, A. and Adda, G.2007. A state-of-the-art statistical machine translation system based on Moses. IN: *Proceedings of MT Summit XI, 10-14 September 2007, Copenhagen, Denmark*, pp.127-133.

Dijkstra, A., Grainder, J., and Van Heuven, W. J. B. 2000a. Recognizing Cognatves and Interlingual Homographs: The Neglected Role of Phonology. *Journal of Memory and Language, 41,* pp. 496-518.

Dijkstra, A., De Brujin, E., Schriefers, H., and Ten Brinke, S. 2000b. More on Interlingual Homograph Recognition: Language Intermixing versus Explicitness of Instruction. *Bilingualism: Language and Cognitive, 3,* pp. 55-64.

Anastasiou, D. 2008. Identification of idioms by machine translation: a hybrid research system vs. three commercial systems. IN: *Proceedings of EAMT 2008: 12th annual conference of the European Association for Machine Translation, September 22 and 23, 2008, Hamburg, Germany*, pp. 12-20.

Doherty, S and O'Brien, S. 2009. Can MT Output be Evaluated Through Eye Tracking? IN: *Proceedings of MT Summit XII: proceedings of the twelfth Machine Translation Summit, August 26-30, 2009, Ottawa, Ontario, Canad*a, pp. 214-221.

Doherty, S., O'Brien, S. and Carl, M. 2010. Eye tracking as an MT evaluation technique. *Machine Translation*, 24, pp. 1-13.

Douglas, S. and Hurst, M. 1996. Controlled language support for Perkins Approved Clear English (PACE). IN: Proceedings of the First International Workshop on Controlled Language Applications, Leuven, Belgium, pp. 93-105.

Dragsted, B. 2004. Segmentation in Translation and Translation Memory Systems: *An Empirical Investigation of Cognitive Segmentation and Effects of Integrating a TM-System into the Translation Process*. PhD thesis: Copenhagen Business School.

Dragsted, B. and Hansen, I. G. 2008. Comprehension and production in translation: a pilot study on segmentation and the coordination of reading and writing processes. IN: Göpferich, S., Jakobsen, A. L. and Mees, I. M. (eds). *Looking at Eyes. Eye-Tracking Studies of Reading and Translation Processing, Copenhagen Studies in Language 36*. Copenhagen : Samfundslitteratur, pp. 9-30.

Drum, P. A., Calfee, R. C., and Cook, L. K. 1981. The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly, 16* (4), pp. 486-514.

Du, J., He, Y., Penkale, S. Way, A. 2009. MaTrEx: the DCU MT System for WMT 2009. IN: *Proceedings of the Fourth Workshop on Statistical Machine Translation, EACL 2009, Athens, Greece.*

DuBay, W. H. 2004, *The Principles of Readability* (Online). Available from: www.nald.ca/fulltext/readab/readab.pdf (Accessed August 15, 2009).

Duchowski, A. 2003. *Eye-Tracking Methodology – Theory and Practice*. London: Springer-Verlag.

Duffy, T. M. 1985. Readability formulas: What's the use? IN: Duffy, T. M., Waller, R. M (eds.). *Designing usable texts*. New York: Academic Press, pp. 113-143.

Dugast, L., Senellart, J. and Koehn, P. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. IN: *Proceedings of the 2nd Workshop on Statistical Machine Translation (WSMT 2007), Prague, Czech Republic*, pp.220-223.

Dugast, L., Senellart, J. Philipp Koehn, P. 2008. Can we relearn an RBMT system? IN: *Proceedings of ACL-08: HLT. Third Workshop on Statistical Machine Translation, June 19, 2008, The Ohio State University, Columbus, Ohio, USA*, pp. 175-178.

d'Ydewalle, G. and de Bruycker, W. 2007. Eye movements of children and adults while reading television subtitles. *European Psychologist*, 12 (3), pp. 196-205.

d'Ydewalle, G., Praet, C., Verfaillie, K., and Van Rensbergen, J. 1991. Watching subtitled television: Automatic reading behaviour. *Communication Research*, 18, pp. 650-666.

Ebbinghaus, H. 1885/1913. *Uber das Gedachtnis* (Leipzig: Dunker) (translated by H. Ruyer and C. E. Bussenius). New York: Teacher College, Columbus University.

Eger, N., Ball, L. J., Stevens, R. and Dodd, J. 2007. Cueing Retrospective Verbal Reports in Usability Testing Through Eye Movement Replay. IN: *Proceedings of BCS HCI*, pp. 129-137.

Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., Hunsicker, S., Herrmann, T. and Chen, Y. 2008. Hybrid machine translation architectures within and beyond the EuroMatrix project. IN: *Proceedings of EAMT 2008: 12th annual conference of the European Association for Machine Translation, September 22 and 23, 2008, Hamburg, Germany*, pp. 27-34.

Elliston, J.S.G. 1979. Computer-aided translation – a business viewpoint. IN: Proceedings of Translating and the Computer 1: proceedings of a seminar, London, 14th November, pp. 149-158.

Entin, E. B., and Klare. G. 1985. The relationships of measures of interest, prior knowledge, and readability to comprehension of expository passages. IN: *Advances in Reading/Language Research, Vol. III. Greenwich, CN: JAI Press.*

Ericsson, K. A. 2000. How experts attain and maintain superior performance: Implications for the enhancement of skilled performance in older individuals. *Journal of Aging and Physical Activity, 8*, pp. 346–352.

Ericson, K. A., and Simon, H. A. 1980. Verbal reports as data. *Psychological Review, 87,* pp. 215-251.

Ericson, K. A., and Simon, H. A. 1984. *Protocol Analysis: Verbal Reports as Data*. Cambridge, Massachusetts: MIT.

Ericson, K. A., and Simon, H. A. 1987. *Protocol Analysis: Verbal Reports as Data (2nd ed.)*. Cambridge, Massachusetts: MIT.

Eysenck, M. W. and Keane, M. T. 2008. *Cognitive Psychology: A Student's Handbook* (5th ed.). East Sussek and New York: Psychology Press.

Eysenck, M. W. and Keane, M. T. 2010. *Cognitive Psychology: A Student's Handbook* (6th ed.). East Sussek and New York: Psychology Press.

FEMTI - A Framework for the Evaluation of Machine Translation in ISLE. (Online). Available from: www.issco.unige.ch:8080/cocoon/femti/st-home.html (Accessed 1 August 2010).

Flanagan, M. 1997. MT Today: Emerging Roles, New Successes. *Machine Translation, 12,* pp. 25-27.

Flanagan, M. 2009. *Recycling text: human evaluation of example-based machine translation subtitles for DVD*. PhD thesis. Dublin City University.

Flesch, R. 1943. *Marks of Readable Style: A Study in Adult Education*. New York: Bureau of Publications, Teachers College, Columbia University.

Forcada, M. L. 2010. Machine translation today. IN: Gambier, Y., and Doorslaer, L. (eds.). *Handbook of Translation Studies,* 1, 215-223. Amsterdam and Philadelphia: John Benjamins.

Frazier, L., and Rayner, K. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14, pp. 178-210.

Frenck-Mestre, C. 2005. Eye-movement recording as a tool for studying syntactic processing in a second language: a review of methodologies and experimental findings. *Second Language Research*, 21 (2), pp. 175-198.

Frey, L. R., Botan, C. H., Friedman, P. G. and Kreps, G. L. 1991. *Investigating Communication. An Introduction to Research Methods*. London: Prentice Hall International.

Fuchs, N. and Schwitter, R. 1996. Attempto Controlled English (ACE). IN: *Proceedings of the First International Workshop on Controlled Language Applications, Katholieke Universiteit Leuven, 26-27 March 1996.*

Gdaniec, C. 1994. The Logos Translatability Index. In Proceedings of the First Conference for Machine Translation in the Americas, Columbia, MD, pp.97–105.

Gerganov, A. 2007. *Eye Tracking Studies with Tobii 1750 - Recommended Settings and Tests* (Online). Available from: http://cogs.nbu.bg/eye-to-it/del/EYE-TO-IT_D1.2_A.pdf (Accessed 19 October 2011).

Gerganov, A., Kaiser, V., Braunstein, V., Popivanov, I., Brunner, C., Neuper, C., and Stamenov, M. 2008. Priming bilingual brain with correct and incongruent translations of true and false cognates in English-German during a translation task: An EEG and eye tracking study. IN: *Ghent Workshop on Bilingualism Ghent, Belgium.*

Gerver, D. 1976. Empirical Studies of Simultaneous Interpretation: A Review and a Model. IN: Brislin, R. (ed.), *Translation: Applications and Research.* New York: Gardner, pp. 165-207.

Gile, D. 1995. *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam and Philadelphia: John Benjamins.

Gile, D. 1998. Observational studies and experimental studies in the investigation of conference interpreting. *Target*, 10 (1), pp. 69-93.

Giles, T. and Still, B. 2005. A Syntactic Approach to Readability. *Journal of Technical Writing and Communication,* 35, pp. 47-70.

Glazner, M., and Cunitz, A. R. 1966. Tactile short-term memory. *Quarterly Journal of Experimental Psychology*, 21, pp. 180-184.

Godden, D. R., and Baddeley, A. D. 1975. Context-dependent Memory in Two Natural Environments: On Land and Underwater. *British Journal of Psychology*, 66, pp. 325-332.

Godden, K. 1998. Controlling the Business Environment for Controlled Language. IN: Mitamura *et al*. (eds), pp. 185-189.

Godden, K. 2000. The Evolution of CASL Controlled Authoring at General Motors. IN: *Proceedings of the Third International Workshop on Controlled Language Applications, CLAW 2000. Seattle, WA,* pp. 14–19.

Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N., and Wichansky, A. M. 2002. Eye tracking in web search tasks: Design implications. IN: *Proceedings of the Eye Tracking Research and Applications Symposium 2002, NY: ACM Press*, pp. 51-58.

Göpferich, S., Jakobsen, A.L. and Mees, I. M. 2008. *Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing. Copenhagen Studies in Language 36*. Kopenhagen: Samfundslitteratur.

Govyaerts, P. 1996. Controlled English, Curse or Blessing? A User's Perspective. IN: *Proceedings of the First Controlled Language Application Workshop (CLAW 1996), Centre for Computational Linguistics, Leuven, Belgium*, pp. 137-142.

Graesser, A., McNamara, D. S., Louwerse, M., & Cai, Z. 2002. Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers, 36,* pp. 193-202.

Gray, W.S. and Leary, B.E. 1935. *What makes a book readable*. Chicago: University of Chicago Press.

Green, T. 1979. The necessity of syntax markers: two experiments with artificial languages. *Journal of Verbal Learning and Behavior,* 18, pp. 481-496.

Greenspoon, J., and Ranyard, R. 1957. Stimulus conditions and retroactive inhibition. *Journal of Experimental Psychology,* 53, pp. 55-59.

Groves, D. and Way, A. 2005. Hybrid example-based SMT: the best of both worlds? IN: *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond, 29-30 June, Michigan, USA,* pp. 183-190.

Groves, Declan. (2007). *Hybrid Data-Driven Models of Machine Translation*. PhD thesis. Dublin City University.

Gunning, R. 1952. *The Technique of Clear Writing.* New York, McGraw-Hill.

Hannu, K., and Pallab, P. 2000. A comparison of concurrent and retrospective verbal protocol analysis. *The American Journal of Psychology*, 113 (3), pp. 387-404.

Hansen, G. 2003. Controlled the process: Theoretical and methodological reflections on research into translation processes. In F. Alves (ed.). *Triangulation*

Translation: Perspectives in Process Oriented Research. Amsterdam and Philadelphia: John Benjamins, pp. 25-42.

Hargis, G. 2000. Readability and computer documentation. *ACM Journal of Computer Documentation*, 24, (3), pp. 122-131.

Harrison, C. 1980. *Readability in the classroom.* Cambridge: Cambridge University Press.

Hassan, H., Hearne, M., Way. A. and Khalil S. 2006. Syntactic Phrase-Based Statistical Machine Translation. IN: *Proceedings of the IEEE 2006 Workshop on Spoken Language Translation, Palm Beach, Aruba.*

Hayes, P. Maxwell, S. and Schmandt, L. 1996. Controlled English Advantages for Translated and Original English Documents. IN: *Proceedings of The First International Workshop On Controlled Language Applications, Katholieke Universiteit Leuyen, Belgium,* pp. 84-92.

Healy, A. F., and McNamara, D. S. 1996. Verbal learning and memory: Does the modal model still word? IN: Spence, J. T., Darley, J. M. and Foss, D. J. (eds.). *Annual Review of Psychology*, 47, pp. 143–172.

Hearne, M., and Way, A. 2011. Statistical Machine Translation: A Guide for Linguists and Translators. IN: *Language and Linguistics Compass* (in press).

Henry, G. 1973. *Une technique de mesure de la lisibilité, specifique de la langue francaise.* PhD thesis. Université de Liege.

Hess, E. H. and Polt, J. M. 1964. Pupil Size in Relation to Mental Activity in Simple Problem Solving. *Science,* 143, pp. 1190-1192.

Hoard, J. E., Wojcik, R. and Holzhauser, K. 1992. An Automated Grammar and Style Checker for Writers of Simplified English. IN: Holt, P., and Williams, N. (eds.). *Computers and Writing: Stage of the Art*. Intellect: UK, pp. 278-296

Hollingworth, A., and Henderson, J.M. 2002. Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance,* 28, pp. 113-136.

Holsanova, J., Rahm, H. and Holmqvist, K. 2006. Entry points and reading paths on newspaper spreads: comparing a semiotic analysis with eye-tracking measurements. *Visual Communication*, 5, pp. 65-93.

Holt, P. and Williams, N. (eds). 1992. *Computers and Writing: State of the Art*. Oxford: Kluwer.

Homan, S., Hewitt, M., and Linder, J. 1994. The development and validation of a formula for measuring single-sentence test item readability. *Journal of Educational Measurement*, 31 (4), pp. 349-358.

Howitt, D. and Cramer, D. 2008. *Introduction to Statistics in Psychology* (4th ed). Essex, England: Pearson Education Limited.

Huang, F., and Papineni, K. 2007. Hierarchical system combination for machine translation. IN: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pp. 277-286.

Huijsen, W. O. 1998, Controlled Language – An Introduction. IN: *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW 98), May 21-22, 1998, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA*, pp. 1-15.

Hulme, C., Maughan, S., and Brown, G. D. A. 1991. Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, 30, pp. 685-701.

Hutchins. J. 1987. Prospects in machine translation. IN: *Proceedings of MT Summit: Machine Translation Summit. Manuscripts and Program, September 17-19, 1987, Hakone Prince Hotel, Japan*, pp. 48-52.

Hutchins, J. 2003. *Machine translation and computer-based translation tools: what's available and how it's used*. University of Valladolid, Spain.

Hutchins, J. 2005. Towards a definition of example-based machine translation. IN: *Proceedings of Second Workshop on Example-Based Machine Translation, Phuket, Thailand, September 16, 2005*, pp. 63-70.

Hutchins, J., and Somers, H. 1992. *An introduction to machine translation*. London: Academic Press.

Hvelplund, K. T. 2011. *Allocation of Cognitive Resources in Translation: An Eye-tracking and Key-logging Study*. PhD thesis. Copenhagen Business School

Hyde, T. S., and Jenkins, J. J. 1969. Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology,* 82, pp. 472-481.

Hyönä, J. and Nurminen, A. M. 2006. Do adult readers know how they read? Evidence from eye movement patterns and verbal reports. *British Journal of Psychology,* 97, pp. 31-50.

Hyönä, J., and Koivisto, M. 2006. The role of eye movements in lateralized word recognition. *Laterality*, 11, pp. 155-169.

Hyönä, J., Tommola, J. and Alaja, A. M. 1995. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *Quarterly Journal of Experimental Psychology*, 143, pp. 1190-1192.

Iqbal, S. T., Adamczyk, D. Zheng, X. and Bailey, B. P. 2005. Towards an index of opportunity: understanding changes in mental workload during task execution. *Proceedings of CHI 2005, 2nd-7th April 2005, Oregon USA*. pp. 311-320.

Isabelle, P. 1987. Machine Translation at the TAUM Group. IN King, M. (ed.). *Machine Translation Today.* Edinburgh University Press, pp. 247-277.

Isham, W. P. 1994. On the Relevance of Signed Languages to Research in Interpretation. *Target, Interpreting Research (15),* pp. 135-149.

Jääskeläinen, R. and Tirkkonen-Condit, S. 1991. Automatised processes in professional vs. non-professional translation: A think-aloud protocol study. IN: Tirkkonen-Condit, S. (ed.). *Empirical Research in Translation and Intercultural Studies. Selected Papers of the TRANSIF Seminar*. London: Continuum, pp. 48-53.

Jääskeläinen, R. 1999. *Tapping the Process: An Explorative Study of the Cognitive and Affective Factors Involved in Translating*. PhD thesis. University of Joensuu.

Jakobsen, A. L. 2002. Translation drafting by professional translators and by translation studies. IN: Sánchez Trigo, E. and Fouces, O. D. (eds). *Traducción and Comunicación* 3. Vigo: Universidade de Vigo, Servicio e Publicacións. pp. 69-95.

Jakobsen, A. L. 2003. Effects of think aloud on translation speed, revision and segmentation. IN: Alves, F. (ed.). *Triangulating Translation. Perspectives in Process Oriented Research*. Amsterdam: John Benjamins, pp. 69-95.

Jakobsen, A. L. 2006. Research methods in translation - Translog. IN: Sullivan, K. P. H. and Lindren, E. (eds.). *Computer Keystroke Logging and Writing*. Amsterdam, Elsevier, pp. 95-105.

Jakobsen, A. L. and Jensen, K. T. H. 2008. Eye movement behaviour across four different types of reading task. IN: Göpferich, S., Jakobsen, A. L. and Mees, I. M. (eds). *Looking at Eyes: Eye-Tracking Studies of Reading and Translation*

Processing, Copenhagen Studies in Language 36. Copenhagen: Samfundslitteratur, pp. 103-124.

Jakobsen, A. L. and Schou, L. 1999. Translog Documentation Version 1.0. IN: Hansen, G. (ed.). *Probing the Process of Translation: Methods and Results Appendix 1. Copenhagen Studies in Language 24.* Copenhagen: Samfundslitteratur, pp. 103-124.

James, W. 1890. *Principles of Psychology*. New York: Holt.

Jensen, K. T. H. 2009. Indicators of text complexity. IN: Mees, I. M., Alves, F. and Göpferich, S. (eds*). Methodology, Technology and Innovation in Translation Process Research: A Tribute to Arnt Lykke Jakobsen, Copenhagen Studies in Language 38.* Copenhagen: Samfundslitteratur.

Jensen, K. T. H., Sjørup, A. C., Balling, L. W. 2009. Effects of L1 syntax on L2 translation. IN: Mees, I. M., Alves, F. and Göpferich, S. (eds*). Methodology, Technology and Innovation in Translation Process Research: A Tribute to Arnt Lykke Jakobsen, Copenhagen Studies in Language 38.* Copenhagen: Samfundslitteratur, pp. 319-336.

Jerabek, I. and Standing, L. 1992. Imagined test situations produce contextual memory enhancement. *Perceptual and Motor Skills* 75, pp. 381-400.

Jones, M. J. 1988. A longitudinal study of the readability of the chairman's narratives in the corporate reports of a UK company. *Accounting and Business Research*, 18 (72), pp. 297-305.

Just, M. A. and Carpenter, P. A. 1980. A theory of reading: from eye fixations to comprehension. *Psychological Review*, 87 (4), pp. 329-354.

Kaakinen , J. K. and Hyönä, J. 2005. Perspective effects on expository text comprehension: Evidence from think-aloud protocols, eye tracking and recall. *Discourse Processes*, 40, pp. 239-257.

Kaakinen, J., Hyönä, J. and Keenan, J. 2003. How prior knowledge, WMC, and relevance of information affect eye fixations in expository text. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 29 (3), pp. 447-457.

Kagdi, H., Yusuf, S., and Maletic, J. I. 2007. On using eye tracking in empirical assessment of software visualizations. IN: *Proceedings of the 1st ACM international Workshop on Empirical Assessment of Software Engineering Languages and Technologies: Held in Conjunction with the 22nd IEEE/ACM international Conference on Automated Software Engineering, Atlanta, Georgia, November 05 - 05, 2007, WEASELTech '07. ACM, New York, NY*, pp. 21-22.

Kahan, T. L., and Johnson, M. K. 1992. Self effects in memory for person information. *Social Cognition*, 10 (1), pp. 30-50.

Kaljurand K. 2008. ACE View — an ontology and rule editor based on Attempto Controlled English. IN: *5th OWL Experiences and Directions Workshop, 26-27 October, Karlsruhe, Germany.*

Kamprath, C., E. Adolphson, T. Mitamura and E. Nyberg. 1998. Controlled Language Multilingual Document Production: Experience with Caterpillar Technical English. IN: *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW98), Pittsburgh, PA*, pp. 51–61.

Kandel, L. and Moles, A. 1958. Application de l'indice de Flesch a la langue francaise. *Cahiers d'Etudes de Radio-Télévision*, 19, pp. 253-274.

Kellogg, R. T. 1996. A model of working memory in writing. IN: Levy, C. M. and Ransdell, S. E. (eds*). The Science of Writing*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 57-71.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London/New York: Longman.

Kenny, D. and Opitz, C. 2000. Process Studies in Translation. *Translation Ireland*, 14(1).

Kintsch, W. 1988. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review*, 95, pp. 163-182.

Klare, G. 1963. *The Measurement of Readability*. Ames: Iowa State Press.

Klare, G. 1974. Assessing readability. *Reading Research Quarterly*, 10, pp. 62-102.

Klare, G. 1976. A second look at the validity of readability formulas. Invited essay. *Journal of Reading Behavior*, 8, pp.129-152.

Klare, G. 1984. *Readability*. IN: Pearson, P.D., Barr, R. Kamil, M.L. and Mosenthal, P. (eds.). *The Handbook of Reading Research.* New York: Longman, pp. 681-731.

Klare, G. 2000. The Measurement of Readability. Ames, IA: Iowa State University Press, 1963. Reprinted in *ACM Journal of Computer Documentation*, 24, pp. 107-121.

Klare, G., Mabry, J. E., and Gustafson, L. M. 1955. The relationship of human interest to immediate retention and to acceptability of technical material. *Journal of Applied Psychology,* 35, pp. 92-95.

Klein, S. B., Loftus, J., and Burton, H. 1989. Two self-reference effects: The importance of distinguishing between self-descriptiveness judgments and autobiographical retrieval in self-referent encoding. *Journal of Personality and Social Psychology*, 56, pp. 853-865.

Knops, U. and Depoortere, B. 1998. Controlled Language and Machine Translation. IN: *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW-98), Pittsburgh, PA.*

Koehn, P. 2004. Pharaoh: A Beam Search Decoder for phrase-based Statistical Machine Translation Models. IN: *Proceedings of AMTA-04, Berlin/Heidelberg, Germany.* Springer Verlag, pp. 115–124.

Koehn, P. 2005. Europarl: a parallel corpus for statistical machine translation. IN: Proceedings of MT Summit X, Phuket, Thailand, September 13-15, 2005, pp. 79-86.

Koehn, P. 2009. *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koolstra, C.M., Van der Voort, T.H.A., and d'Ydewalle, G. 1999. Lengthening the presentation time of subtitles on television: Effects on children's reading time and recognition. *Communications*, 24, pp. 407-422.

Koehn, P., Och, F. and Marcu, D. 2003. Statistical phrase-based translation. IN: Proceedings of HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics, May 27 – June 1, 2003, Edmonton, Canada, pp. 48-54.

Kulesza, A. and Shieber, S. 2004. A learning approach to improving sentence-level MT evaluation. IN: *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007), Skövde, Sweden.*

Kwolek, E. 1973. A readability survey of technical and popular literature. *Journalism Quarterly*, 50, pp. 255-264.

Lagarda, A. L., Alabau, V., Casacuberta, R., Silva, R. and Díaz de Ilarraza, E. 2009. Statistical post-editing of a rule-based machine translation system. IN: *Proceedings of the NAACL HLT 2009: Short Papers, Boulder, Colorado.* pp. 217-220.

Lashley, K. S. 1958. Cerebral organization and behavior. IN: Solomon, H. C., Cobb, S. and Penfield, W. (eds.). *The Brain and Human Behavior,* 36. Association for Research in Nervous and Mental Diseases, Research Publications: Williams and Wilkins, pp. 1-18.

Lehrberger, J. and Bourbeau, L. 1988. Machine Translation: Linguistic characteristics of MT Systems and general methodology of evaluation. *Linguistic Investigations: Supplemental*, 15. Amsterdam. John Benjamins.

Leijten, M. and Van Waes, L. 2005. *Inputlog: a logging tool for the research of writing processes.* Antwerpen: Universitat Antwerpen.

Lewis, N.R., Parker, L.D., Pound G.D., and Sutcliffe P. 1986, Accounting report readability: the use of readability techniques. *Accounting and Business Research*, Summer, pp. 199-213.

Lively, B.A. and Pressey, S. L. 1923. A method for measuring the 'vocabulary burden' of textbooks. *Educational Administration and Supervision*, 9, (7), pp. 389-98.

Lorge, I. 1939. Predicting reading difficulty of selections for children. *Elementary English Review,* 16, pp. 229-233.

Luck, S. J., and Vogel, E. K. 1997. The capacity of visual working memory for features and conjunctions. *Nature*, 390, pp. 279-281.

Lux, V. and Dauphin, E. 1996. Corpus Studies: a Contribution to the Definition of a Controlled Language. IN: *Proceedings of the First International Workshop on Controlled Language Applications (CLAW 96), Leuven, Belgium*, pp. 193–204.

McConkie, G. W., and Rayner, K. 1975. The span of the effective stimulus during a fixation in reading. *Perception and Psychophysics*, 17, pp. 578-586.

McCord, M., and Bernth, A. 1998. The LMT Transformational System: Machine translation and the information soup. IN: Proceedings of the third conference of the Association for Machine Translation in the Americas, AMTA '98, Langhorne, PA, USA, October 1998, pp. 344-355.

McLaughlin G. H.1966. *What Makes Prose Understandable*. PhD thesis, University College, London.

Means, L. and Godden, K. 1996. The Controlled Automotive Service Language (CASL) Project. IN: *CLAW 96: Proceedings of the First International Workshop on Controlled Language Applications, Leuven, Belgium*, pp. 106–114.

Michael, N., Johns, M., Owen, C., and Patterson, J. 2008. Effects of caffeine on alertness as measured by infrared reflectance oculography. *Psychopharmacology*, 200, pp. 255-260.

Miller, G. A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63, pp. 81-97.

Mitamura, T. and Nyberg, E. 1995. Controlled English for Knowledge-Based MT: Experience with the KANT System. IN: *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 95), Leuven, Belgium.*

Mitamura, T., Nyberg, E., Adriaens,G., Schmandt, L., Wojcik, R. and Zajac, R. (eds.). 1998. *Proceedings of the Second International Workshop on Controlled Language*

Applications (CLAW 98). Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Moravcsik, J. E., and Kintsch, W. 1993. Writing quality, reading skills, and domain knowledge as factors in text comprehension. *Canadian Journal of Experimental Psychology,* 4 (7), pp. 360-374.

Morrissey, S. 2008. *Data-Driven Machine Translation for Sign Languages*. PhD thesis. Dublin City University.

Mossop, B. 2003. *An Alternative to Deverbalization* (Online) Available from: http://www.yorku.ca/brmossop/Deverbalization.htm (Accessed 1 August 2011).

Nagao, M. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. IN: Elithorn, A., and Banerji, R. (eds). *Artifical and human intelligence: edited review papers presented at the international NATO symposium, October 1981, Lyon, France*. Amsterdam: North Holland, pp. 173-180.

Nakayama, M., Takahashi, K., and Shimizu, Y. 2002. The Act of Task Difficulty and Eye-movement Frequency for the 'Oculo-motor indices'. IN: *Eye Tracking Research and Applications (ETRA) Symposium, ACM*, pp. 43-51.

Neely, J. H. 1991. Semantic Priming Effects in Visual Word Recognition: A Selective Review of Current Findings and Theories. IN: Besner, D., and Humphreys, G. W. (eds.), *Basic Processes in Reading.* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 264-336.

Newell, A. and Simon, H. A. 1972. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nickerson, R. A., and Adams, M. J. 1979. Long–term memory for a common object. *Cognitive Psychology*, 11, pp. 287–307.

Nießen. S and Ney, H. 2001. Toward hierarchical models for statistical machine translation of inflected languages. IN: ACL-EACL 2001 workshop on data-driven machine translation, July 7, 2001, Toulouse, France, pp. 47-54.

Nirenburg, S. (ed.) 1987. *Machine translation: theoretical and methodological issues.* Cambridge, Cambridge University Press.

Nizar H.,and Sadat, F. 2006. Arabic preprocessing schemes for statistical machine translation. IN: *HLT-NAACL 2006: Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, New York, NY, USA, June 2006*, pp. 49-52.

Nyberg, E. and Mitamura, T. 1996. Controlled Language and Knowledge-Based Machine Translation: Principles and Practice. IN: *Proceedings of the First Controlled Language Application Workshop (CLAW 1996), Leuven, Belgium, Centre for Computational Linguistics,* pp. 74-83.

Nyberg, E., Mitamura, T. and Huijsen, W.O. 2003. Controlled language for authoring and translation. IN: Somers, H. (ed.). *Computers and translation: a translator's guide*. Amsterdam: John Benjamins, pp. 245-281.

O'Brien, S. 2003. Controlling Controlled English: An Analysis of Several Controlled Language Rule Sets. IN: *Proceedings of EAMT-CLAW 03, Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Controlled Translation, Dublin, Ireland,* pp. 105-114.

O'Brien, S. 2006. Eye-Tracking and Translation Memory Matches. Perspectives: *Studies in Translatology,* 14 (3), pp. 185-205.

O'Brien, S. 2006. *Machine-translatability and post-editing effort : an empirical study using Translog and choice network analysis*. PhD thesis. Dublin City University.

O'Brien, S. 2008. Processing Fuzzy Matches in Translation Memory Tools – an Eye-tracking Analysis. IN: Göpferich, S., Jakobsen, A. L. and Mees, A. (eds). *Looking at eyes – Eye Tracking Studies of Reading and Translation Processing. Copenhagen Studies in Language 36.* Copenhagen: Samfundslitteratur, pp. 79-102.

O'Brien, S. and Roturier, J. 2007. How portable are controlled languages rules: a comparison of two empirical MT studies. IN: *Proceedings of the 11th Machine Translation Summit of the International Association for Machine Translation (MT Summit XI), Copenhagen, Denmark.*

O'Brien, S. 2009. Controlled language and readability. IN: Shreve, G. and Angelone, E (eds). *Translation and Cognition*. American Translators Association Scholarly Monograph Series, John Benjamins.

O'Brien, S. 2010. Eye tracking in translation process research: methodological challenges and solutions. *Copenhagen Studies in Language, 38,* pp. 251-266.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., and Radev, D. 2004. A smorgasbrod of features for statistical machine translation. IN: *Proceedings of HLT-NAACL 2004: Human Language Technology conference and North American Chapter of the Association for Computational Linguistics annual meeting, May 2-7, 2004, The Park Plaza Hotel, Boston, USA*, pp. 161-168.

Och, F. J. and Ney, H. 2004. Minimum error rate training for statistical machine translation. IN: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan,* pp. 160-167.

Och, F., Tillmann, C. and Ney, H. 1999. Improved alignment models for statistical machine translation. IN: P*roceedings of joint SIGDAT conference on Empirical Methods in Natural Language Processing and Very Large Corpora, University of Maryland, College Park, MD, USA,* pp. 20-28.

Ogden, C. K. 1930. *Basic English: A General Introduction with Rules and Grammar*. London: Paul Treber.

Olive, T. 2004. Working memory in writing: empirical evidence from the dual-task technique. *European Psychologist*, 9, pp. 32-42.

Oppenheim, A. M. 1966. *Questionnaire Design and Attitude Measurement.* New York: Basic Books.

Padilla, P., Bajo, M. T. and Pedilla, F. 1999. Proposal for a cognitive theory of translation and interpreting: a methodology for future empirical research*. The Interpreter's Newsletter*, 9, pp. 61-78.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. 2002. BLEU: A method for automatic evaluation of machine translation. IN: *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL 2002), Philadelphia, Pennsylvania*, pp. 311-318.

Parkin, A. J. 1983. The relationship between orienting tasks and the structure of memory traces - evidence from false recognition. *British Journal of Psychology,* 74, pp. 61-69.

Parton, K., and McKeown, K. 2010. MT error detection for cross-lingual question answering. IN: Proceedings of Coling 2010: 23rd International Conference on Computational Linguistics, 23-27 August 2010, Beijing International Convention Center, Beijing, China, pp. 946-954.

Pascual-Leone, J. 1970. A mathematical model for the transition rule in Piaget's developmental stages. *Acta Psychologica*, 63, pp. 301–345.

Pavlović, M., and Burchardt, A. 2011. From human to automatic error classification for machine translation output. IN: *Proceedings of the 15th*

conference of the European Association for Machine Translation, 30-31 May 2011, Leuven, Belgium, pp. 265-272.

Pavlović, N. and Jensen, K. T. H. 2009. Eye tracking translation directionality. IN: Pym, A. and Perekrestenko, A (eds*). Translation Research Projects 2.* Tarragona: Universitate Rovira i Virgili. pp. 101-119.

Perego, E. and Ghia, E. 2011. Subtitle consumption according to eye tracking data. An acquisitional perspective. IN: Incalcaterra McLoughlin, L. and Miscio, M. (eds.), *Audiovisual and Translation. Theoretical Issues and Didactic Applications.* Peter Lang, Bern, pp. 177-196.

Pickering, M. J., and Traxler, M. J. 1998. Plausibility and recovery from garden paths: An eye tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, pp. 940-961.

Popović, M., and Burchardt, A. 2011. From human to automatic error classification for machine translation output. IN: *The Proceedings of European Association for Machine Translation, Leuven, Belgium,* pp. 265-272.

Power, R., Scott, D. and Hartley, A. 2003. Multilingual Generation of Controlled Languages. IN: *Proceedings of EAMT-CLAW 03, Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Controlled Translation, Proceedings, Dublin, Ireland*, pp. 115–123.

Rapley, T. 2004. Interviews. IN: Seale, C., Gobo, G., Gubrium, J., and Silverman, D. (eds.) *Qualitative Research Practice*. London: Safe Publications.

Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, pp. 372-422.

Rayner, K., Foorman, B.F., Perfetti, C.A., Pesetsky, D., and Seidenberg, M.S. 2001. How psychological science informs the teaching of reading. *Psychological Science in the Public Interest. 2* (2), pp. 31-74.

Rayner, K. and Pollatsek, A. 1989. The Psychology of Reading. Englewood Cliffs, NJ: Prentice-Hall.

Rayner, K. and Sereno, S. 1994: Eye movements in reading: psycholinguistic studies. IN: Gernsbacher, M.A. (ed.). *Handbook of Psycholinguistics.* New York: Academic Press, pp. 57–81.

Read, J. 2000. *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Redish, J. 2000. Readability formulas have even more limitations than Klare discusses. *Journal of Computer Documentation*, 24 (3), pp. 132–137.

Reuther, U. 2003. Two in One – Can it Work? Readability and Translatability by means of Controlled Language. IN: Proceeding of EAMT/CLAW2003, Dublin, Ireland, 15-17.

Richaudeau, F., and Staats, D. M. 1981. Some French work on prose readability and syntax. *Journal of Reading*, 24 (6), pp. 503-508.

Rosenthal, D. M. 2000. Metacognition and Higher-Order Thoughts. *Consciousness and Cognition* 9, pp. 231-242.

Roturier, J. 2004. Assessing a set of controlled language rules: Can they improve the performance of commercial machine translation systems? IN: The *Proceedings of the 26th Conference of Translating and the Computer 26,* London: ASLIB, pp. 1-14.

Roturier, J. 2006. *An Investigation Into the Impact of Controlled English Rules on the Comprehensibility, Usefulness, and Acceptability of Machine-Translated*

Technical Documentation for French and German Users. PhD thesis. Dublin City University.

Roturier, J. 2009. Deploying novel MT technology to raise the bar for quality: A review of key advantages and challenges. IN: *Proceedings of the 12th Machine Translation Summit (MTS 2009), Ottawa, Canada.*

Rudmann, D. S., McConkie, G. W., and Zheng, X. S. 2003. Speech and Gaze: Eye-tracking in cognitive state detection for HCI. In: *Proceedings of ICMI '03.*

Ruiz, C., Paredes, N., Macizo, P., Bajo, M. T. 2008. Activation of lexical and syntactic target language properties in translation*. Acta Psychologica,* 128 (3), pp. 490-500.

Rychtyckyj, N. 2002. An Assessment of Machine Translation for Vehicle Assembly Process Planning at Ford Motor Company. IN: S. Richardson (ed.) *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002, Tiburon, CA, USA, October 8-12. LNAI 2499, Springer-Verlag: Berlin Heideiberg.* pp. 207-215.

Rychtyckyj, N. 2006. Machine Translation for Manufacturing: A Case Study at Ford Motor Company. IN: *Proceedings of the 18th Innovative Applications of Artificial Intelligence Conference (IAAI-2006), Boston, MA, July 18-20, 2006.*

Schachtl, S. 1996. Requirements for Controlled German in Industrial Applications. IN: *Proceedings of the First International Workshop on Controlled Language Applications (CLAW 96), Leuven, Belgium,* pp. 143–149.

Schäler, R., Way. A and Carl., M. 2003. Example-Based Machine Translation in a Controlled Environment. IN: Carl, M. and Way, A. (eds.). *Recent Advances in Example-Based Machine Translation*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 83–114.

Schallert, D. L., and Reed, J. H. 1997. The pull of the text and the process of involvement in one's reading. IN: Guthrie, J. T. and Wigfield, A. (eds.). *Reading engagement: Motivating readers through integrated instruction*. Newark, DE: International Reading Association, pp. 68-85.

Schneider, W., Korkel, J., and Weinert, F. E. 1990. Expert knowledge, general abilities, and text processing. IN: Schneider, W., and Weinert, F. E. (eds.). *Interactions Among Aptitudes, Strategies, and Knowledge in Cognitive Performance*. Springer-Verlag: New York, pp. 235-251.

Schriver, K. A. 1989. Evaluating text quality: The continuum from text-focused to reader-focused methods. IEEE Trans. *Professional Communication*, 32, pp. 239–255.

Schriver, K. A. 2000. Readability formulas in the new millennium: What's the use? *Journal of Computer Documentation,* 24 (3), pp. 138–140.

Schultheis, H. and Jameson, A. 2004. Assessing cognitive load in adaptive Hypermedia Systems: physiological and behavioural methods. IN: *Adaptive Hypermedia and Adaptive Web-Based Systems, Lecture Notes in Computer Science Vol. 3137, Berlin-Heidelberg: Springer*, pp. 225-234.

Scoville, W. B. and Milner, B. 1957. Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurological Psychology,* 20, pp. 11–21.

Seleskovitch, D. 1976. Interpretation: a psychological approach to translating. IN: Brislin, R. W. (ed.). *Translation: Applications and Research.* New York: Gardner, pp. 92-116.

Senez, D. 1998. The Machine Translation Help Desk and the Post-Editing Service. IN: *Terminologie and Traduction*, 1, pp. 289-295.

Sharmin, S., Špakov, O., Räihä, K., Jakobsen, A. L. 2008. Where on the screen do translation students look when translating, and for how long? IN: Göpferich, S., Jakobsen, A. L. and Mees, I. M. (eds*). Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing. Copenhagen Studies in Language 36.* Copenhagen: Samfundslitteratur, pp. 31-51.

Shnayer, S. W. 1969. Relationship between reading interest and reading comprehension. IN: Allen J. (ed.) *Reading and Realism*. Newark, Delaware, I.R.A., pp. 698-702.

Shreve, G. M. and Koby, G. S. 1997. What's in the Black Box? Cognitive Science and Translation Studies. IN: Danks, H. J., Shreve, G. M., Fountain, F. B., McBeath, M. K. (eds.). *Cognitive Processing in Translation and Interpreting*. Thousand Oaks, CA: Sage, pp. xi-xviii.

Simard, M., Ueffing, N., Isabelle, P. and Kuhn, R. 2007. Rule-based translation with statistical phrase-based post-editing. IN: *Proceedings of WMT07, Prague, Czech Republic, June. Association for Computational Linguistics,* pp. 203-206.

Slamecka, N. J. and McElree, B. 1983. Normal forgetting of verbal lists as a function of their degree of learning. *Journal of Experimental Psychology. Learning, Memory and Cognition,* 9, pp. 384-397.

Smith, M. and Taffler, R. J. 1992. Readability and Understandability: Different Measures of the Textual Complexity of Accounting Narrative. *Accounting, Auditing, and Accountability Journal,* 5 (4), pp. 84-98.

Smith, R. W. and Healy, A. F. 1998. The time-course of the generation effect. *Memory and Cognition*, 26 (1), pp. 135-142.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. 2006. A study of translation edit rate with targeted human annotation. IN: *Proceedings of the 7th*

Conference of the Association for Machine Translation in the Americas (AMTA 2006), Cambridge, Massachusetts.

Snover, M., Madnani, N., Dorr, B.J. and Schwartz, R. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tuneable MT metric. IN: *Proceedings of the EACL-2009 Workshop on Statistical Machine Translation (WMT09), 30-31 March, Athens, Greece,* pp. 259-268.

Somers, H. 1997. The current state of machine translation. IN: *Proceedings of MT Summit, 29 October – 1 November 1997, San Diego, California, USA*, pp. 115-124.

Specia, L., Cancedda, N. and Dymetman, M. 2010. A dataset for assessing machine translation evaluation metrics. IN: *Proceedings of the seventh international conference on Language Resources and Evaluation, 17-23 May 2010, Valletta, Malta*, pp. 3375-3378.

Sperling, G. 1960. The information available for brief visual presentations. *Psychology Monographs, 75, 11,* pp. 1-29.

Spyridakis, J., Holmback, H., and Shubert, S. K. 1997. Measuring the Translatability of Simplified English in Procedural Documents. *IEEE Transactions on Professional Communication,* 40, (1), pp. 4-12.

Staub, A., Rayner, K., Pollatsek, A., Hyönä, J. and Malewksi, H. 2007. The Time Course of Plausibility Effects on Eye Movements in Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33 (6), pp. 1161-1169.

Stitzlein, C. A., Li, J., and Krumm-Heller, A. 2006. Gaze analysis in a remote collaborative setting. IN: *Proceedings of the 20th Conference of the Computer-Human interaction Special Interest Group (Chisig) of Australia on Computer-Human interaction: Design: Activities, Artefacts and Environments, Sydney, Australia, November 20 - 24, 2006*, pp. 417-420.

Stroppa, N. and A. Way. 2006. MaTrEx: DCU Machine Translation System for IWSLT 2006. IN: *Proceedings of the International Workshop on Spoken Language Translation, Kyoto, Japan*, pp. 31-36.

Sun, Y. 2010. *An Investigation into Automatic Translation of Prepositions in IT Technical Documentation from English to Chinese.* PhD thesis. Dublin City University.

Tashakkori, A. and Teddlie, C. 2003. *Handbook of Mixed Methods in Social and. Behavioral Research.* Thousand Oaks: Sage.

Tatsumi, M. 2009. Correlation between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors. IN*: Proceedings of MT Summit XII, Ottawa, Ontario, Canada*, pp. 332-339.

Tatsumi, M. 2010. *Post-editing Machine Translated Text in a Commercial Setting: Observation and Statistical Analysis*. PhD thesis. Dublin City University.

Taylor, W.L. 1953. Cloze Procedure: A New Tool for Measuring Readability. *Journalism Quarterly,* 30, pp. 415-433.

Terumasa, E. 2007. Rule-based machine translation combined with statistical post editor for Japanese to English patent translation. IN: *Proceedings of MT Summit XI Workshop on patent translation, 11 September 2007, Copenhagen, Denmark,* pp. 13-18.

Tharp, J. B. 1939. The measurement of vocabulary difficulty. *Modern Language Journal*, 24, pp. 169-178.

Thorndike, E.L. 1921. *The Teacher's Word Book*. New York: Teacher's College, Columbia University.

Toury, G. 1985. A rationale for descriptive translation studies. IN: Hermans, T (ed.). *The Manipulation of Literature: Studies in Literary Translation.* New York: St Martins Press.

Trujillo, Arturo. 1999. *Translation engines: Techniques for Machine Translation.* London: Springer.

Tulving, E. 1972. Episodic and semantic memory. IN: Tulving, E. and Donaldson, W. (eds). *Organization of Memory*. London: Academic Press.

Turian, J. P., Shen, L. and Melamed, I. D. 2003. Evaluation of Machine Translation and its Evaluation. IN: *Proceedings of MT Summit IX, New Orleans, USA,* pp. 386-393.

Unwalla, M. 2004. AECMA Simplified English. *Communicator*, Winter Edition, no page numbers.

Valdés, B., Catena, A. and Marí-Beffa, P. 2005. Automatic and controlled semantic processing: a masked prime-task effect. *Consciousness and Cognition*, 14, pp. 278-295.

Van der Eijk, P., M. de Konig and G. van der Steen. 1996. Controlled language correction and translation. IN: *Proceedings of the First International Workshop on Controlled Language Applications, Leuven, Belgium*, pp. 64-73.

Van Gog, T, Kester, L, Nievelstein, F, Giesbers, B and Paas, F. 2009. Uncovering Cognitive Processes: Different Techniques That Can Contribute to Cognitive Load Research and Instruction. *Computers in Human Behavior*, 25, pp. 325-331.

Van Hell, J. G., and Dijkstra, T. 2002. Foreign Language Knowledge Can Influence Native Language Performance in Exclusively Native Contexts. *Psychology Review Bulletin 9,* pp. 780-789.

Van Slype, G. 1979. *Critical methods for evaluating the quality of machine translation.* Final Report, Bureau Marcel van Dijk / European Commission,

Brussels. Available from: http://www.issco.unige.ch/en/research/projects/isle/van-slype.pdf (Accessed December 21, 2011).

Vassiliou, M., Markantonatou, S., Maistros, Y. and Karkaletsis, V.2003. Evaluating Specifications for Controlled Greek. IN: *Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Controlled Translation, Proceedings, Dublin, Ireland*, pp.185-191.

Vilar, D., Xu, J., D'Haro, L.F. and Ney, H. 2006. Error analysis of statistical machine translation output. IN: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006), Genoa, Italy.*

Vogel, M. and Washburne, C. 1928. An objective method of determining grade placement of children's reading material. *Elementary School Journal,* 28, pp. 373-381.

Way, A. and Gough, N. 2005. Controlled translation in an example-based environment: What do automatic evaluation metrics tell us? *Machine Translation*, 19, pp.1-36.

Way, A. and Gough, N. 2003. Controlled Generation in Example-Based Machine Translation. IN: Gough, N. (ed.). *Proceedings of MT Summit IX, New Orleans, LO.,* pp.133-140.

Wilks, Y. 2009. *Machine translation: Its scope and limitations.* New York. Springer.

Williams, C.B. 1940. A note on the statistical analysis of sentence length as a criterion of literary style. *Biometrika,* 31, pp. 356-361.

Yamada, S., Sumita, E. and Kashioka, H. 2000. Translation Using Information on Dialogue Participants. IN: *Proceedings of the 6th Applied Natural Language*

Conference and 1st Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, WA., pp.37–43.

Zhu, Z., Fujimura, K. and Ji, Q. 2002. Real-time eye detection and tracking under various light conditions. IN: *Proceedings of the 2002 symposium on Eye tracking research and applications, ACM Press*, pp. 139-144.

Appendices:

Contents

A. Informed Consent Form

DUBLIN CITY UNIVERSITY

I. Research Study Title

Readability and Comprehensibility of Machine Translation Output

Stephen Doherty, Centre for Next Generation Localisation, Dublin City University

II. Clarification of the purpose of the research

(1) *The purpose of the research is to investigate the readability and comprehensibility of machine translation output as measured via an eye tracker.*

III. Confirmation of particular requirements as highlighted in the Plain Language Statement

You will be asked to attend one session which should last no more than 45 minutes. In this session you will be asked to read five paragraphs that will be displayed on an eye tracking monitor - the eye tracking monitor looks just like a normal computer screen and works in a completely noninvasive way; much like a digital camera. After reading, you will be asked to answer some short questions which will test your level of recall and understanding of the material you have just read. Lastly, you will be asked to rate the texts you read.

Participant – please complete the following (Circle Yes or No for each question)	
Have you read or had read to you the Plain Language Statement	*Yes/No*
Do you understand the information provided?	*Yes/No*
Have you had an opportunity to ask questions and discuss this study?	*Yes/No*
Have you received satisfactory answers to all your questions?	*Yes/No*

IV. Confirmation that involvement in the Research Study is voluntary

Involvement in this study is voluntary. You may withdraw from the Research Study at any point. There will be no penalty for withdrawing before all stages of the Research Study have been completed. Involvement /non-involvement in this study will not affect your relationship with DCU in any way.

V. Advice as to arrangements to be made to protect confidentiality of data, including that confidentiality of information provided is subject to legal limitations

Your anonymity will be protected at all times. You will be given an identifier such as "Participant A" and no link will ever be made to your real identity. The data collated will be used only by Stephen Doherty and will not be given to anybody else.

VII. Signature:

I have read and understood the information in this form. My questions and concerns have been answered by the researchers, and I have a copy of this consent form. Therefore, I consent to take part in this research project

Participants Signature: _____

Name in Block Capitals: _____

Witness: _____

Date: _____

B. Participant Questionnaire

Participant Questionnaire

1. Name _____

2. Occupation _____

3. Educational Background (e.g Business, Science…)

_____

4. Do you have any knowledge of, or experience with working with linguistics or translating?

Yes *No*

If so, please give details:

_____

5. How much time, on average, would you spend using a computer per week?

0 – 10 hours *10 – 20 hours* *20+ hours*

6. Do you have any specific knowledge or experience with anti-virus or data protection/security software?

Yes *No*

C. Uncontrolled Output

A propos de Symantec AntiVirus

Vous pouvez installer Symantec AntiVirus à la protection contre les virus et les risques de sécurité de l'une ou l'autre autonome ou une administrator-managed l'installation. Autonome signifie que votre installation du logiciel Symantec AntiVirus n'est pas géré par un administrateur réseau. Si vous gérer votre propre ordinateur, il doit être l'un des types suivants:

- Ordinateur autonome non connecté à un réseau, tel qu'un ordinateur d'accueil ou un portable autonome, avec l'installation de Symantec AntiVirus utilisant l'une ou l'autre de l'option par défaut paramètres ou de paramètres des options de administrator-preset
- Ordinateur distant qui sur votre réseau d'entreprise respectent qui doit exigences de sécurité avant la connexion

Le paramètre par défaut pour Symantec AntiVirus fournit une protection contre les virus et risques de sécurité pour votre ordinateur. Toutefois, vous pouvez créés resuit jamais besoins de votre société, à optimize les performances système et de désactiver les options qui ne s'applique pas.

Si votre installation est géré par l'administrateur, certaines options peuvent être verrouillé ou non disponibles, ni risque de ne pas apparaître du tout, en fonction de votre politique de sécurité de son apportées par l'administrateur. Votre administrateur s'exécute analyses sur votre ordinateur et peut définir les analyses planifiées. Votre administrateur peut advise vous en tant que vous devez effectuer des tâches pour les éléments à l'aide de Symantec AntiVirus.

Le support technique

Dans le cadre de Symantec Security Response, Symantec Global Support technique gère la prise en charge du groupe de centers pendant toute la planète. Le support technique de son groupe rôle essentiel de répondre aux questions spécifiques de produit fonctionnalité / fonctionner, l'installation et la configuration, ainsi que pour author pour de contenu de notre base de connaissances web-accessible. Le support technique Fonctionnement du groupe collaboratively avec des zones vers d'autres functional dans Symantec answer votre questions dans une timely fashion.

Par exemple, le groupe de support technique fonctionnement engineering du produit, ainsi que Symantec Security Response pour fournir des services d'alerte et des mises à jour des définitions de virus épidémies de virus et les alertes de sécurité. Le support technique de Symantec offerings sont les suivantes:

- Une plage de prendre en charge les options que vous donnez souplesse de droite pour sélectionner la quantité de service de tout société la taille.
- Telephone Web composants qui prennent en charge et fournit une réponse rapide et des informations up-to-the-minute
- Mise à niveau jour que delivers mise à niveau des logiciels de protection automatique
- Les mises à jour de contenu concernant les définitions de virus et les signatures de sécurité s'assurer que le plus haut niveau de protection
- Global Support de Symantec Security Response experts, disponible 24 heures un jour et 7 jours ouvrables dans un worldwide contiennent des langues pour ces enrolled pour prennent en charge les clients du programme

Règles de cheval de Troie

Les chevaux de Troie malveillants sont des programmes qui sont déguisées en utile pour les programmes. Lorsque vous installez et exécutez un cheval de Troie, il apparaît pour être effectue une fonction utile alors mais il n'est jamais damaging votre ordinateur de système d'exploitation. Symantec Client Firewall Règles de cheval de Troie examiner les communications réseau des clients Symantec Client Firewall qui accèdent à Internet, signe en recherchant de ces programmes dommageables. Si elle est détectée, elle opération immédiate à ce type de menace. Règles de cheval de Troie planifier la priorité inférieur à générales ou des règles de programme. Elles sont appliques seulement une fois que ces deux groupes de règles sont appliquées. Règles de cheval de Troie par défaut sont toujours le blocage des règles, dans contrast pour générales ou des règles de programme, qui peuvent autoriser l'accès. Règles de cheval de Troie répar une correspondance de modèles d'attaque associé à la liste des menaces connues contre les communications réseau en permanence. Occasionnellement, inoffensif l'activité du réseau peuvent déclencher un cheval de Troie Alert, si la communication implique d'utiliser des ports spécifiques ou d'autres connus critères associe à un cheval de Troie.

Si vous continually reçoivent les mêmes d'alerte de cheval de Troie, vous pouvez dé investigate dé pour imposer l'alerte n'a pas à génère par une activité normale ou les communications sur votre réseau.

Suppression de fichiers infecté s par des virus en quarantaine

Si vous supprimez un fichier dans la mise en quarantaine, Symantec AntiVirus Supprime définitivement de votre ordinateur actuels jamais disque dur. La suppression d'un fichier infecte par un virus réduit le risque qu'aucun virus peut se propager en supprimant le fichier (et c'est pourquoi le virus) de l'ordinateur. Suppression de le fichier infecté est utile pour les virus de fichier ou de macro. Comme les virus peuvent endommager la suppression des fragments un fichier, le fichier infecté et il remplacement par une copie de sauvegarde Nettoyer le fichier peut être Nettoyage mieux que le fichier infecté. Vous pouvez effectuer cette action après manuellement un fichier infecte n'a à été déplaces dans la quarantaine. Suppression de le fichier infecté dans la zone de quarantaine serait une manière utiles pour supprimer un virus d'un fichier disposable ayant été unable pour être nettoyé. Utilisez cette option uniquement si vous avez Nettoyer sauvegardes de fichiers que vous decided jamais à analyser. N'utilisez pas cette action comme action principale pour les fichiers soumis à Auto-Protect ou les analyses planifiées.

Activation et désactivation d 'Auto-Protect

Si vous n'avez pas modifié les paramètres d'option par défaut, Auto-Protect se charge au démarrage de l'ordinateur pour vous protéger contre les virus et les risques de sécurité. Vérifie les programmes pour rechercher les virus et les risques de sécurité car elles s'exécutent et contrôle de tout ordinateur activité peut indiquer la présence d'un virus ou d'un risque de sécurité. Lorsqu'un virus, les activités suspectes (un événement qui ne peuvent être la présence d'un virus) ou un risque de sécurité est détecté, Auto-Protect vous alerte. Dans certains cas, Auto-Protect peut Avertir d'un concernant les activités suspectes qui vous savez qu'il n'est pas la présence d'un virus. Par exemple, cela peut se produire si vous installez un nouveau programme sur l'ordinateur. Si vous ne sera tel effectue une activité et créé viter l'avertissement, vous pouvez désactiver temporairement Auto-Protect. Veillez à l'activez lorsque vous avez terminé votre tâche pour garantir que votre ordinateur reste protégés. Votre administrateur peut verrouiller Auto-Protect pour que vous ne pouvez pas désactiver

pour raison quelconque, ou spécifier que Auto-Protect pour le système de fichiers peuvent être désactivé, mais reenables temporairement automatiquement après une durée spécifiée.

A propos des inclusions et des exclusions lors des analyses

Des inclusions et des exclusions vous aider à balance la quantité de protection que votre réseau nécessite avec la durée et requis pour fournir des ressources cette protection. Par exemple, si vous choisissez d'analyser tous les types de fichier, vous pouvez décider d'exclure certains dossiers contenant uniquement des fichiers de données qui ne peuvent pas être infectés. Ou, il peut être utile de n'analyser que les fichiers portant des extensions qui sont susceptibles de contenir un virus ou un risque de sécurité. Lorsque vous sélectionnez pour n'analyser que certaines extensions, vous excluez automatiquement tous les fichiers qui portent d'autres extensions de l'analyse. Ces choix diminue le overhead associé à la recherche des fichiers. Selon le type d'analyse et les objets de l'analyse, vous pouvez exclure par fichier, dossier ou type de fichier types de fichier. Vous pouvez inclure seulement certaines types de fichier ou des extensions dans une analyse. Vous pouvez inclure et exclure des à éléments des analyses lancées depuis Symantec Client Security de l'interface utilisateur client ou serveur ou depuis la console Symantec System Center.

D. Controlled Output

À propos de Symantec AntiVirus

Vous pouvez installer Symantec AntiVirus™ de virus et de risque de sécurité en tant que la protection autonomes ou une installation gérée par l'administrateur. Une installation autonome signifie qu'un administrateur réseau n'est pas gérer vos Symantec AntiVirus. Si vous gérez vos propres ordinateur, il doit être un des types suivants:

- Un ordinateur autonome qui n'est pas connecté à un réseau par une installation de Symantec AntiVirus qui utilise les paramètres par défaut administrator-preset ou options
- Un ordinateur distant qui se connecte à votre réseau d'entreprise qui doivent répondre aux spécifications de sécurité avant connexion.

Les paramètres par défaut de Symantec AntiVirus assurent la protection de virus et de risque de sécurité pour votre ordinateur. Cependant, vous pouvez régler adapter aux deux pour votre entreprise doit optimiser les performances du système et désactiver les options qui ne s'appliquent pas. Si votre administrateur gère votre installation, quelques options peuvent être verrouillées ou indisponibles ou ne s'affiche pas à tout, selon la votre administrateur politique de sécurité. Votre administrateur exécute des analyses sur votre ordinateur et peut configurer des analyses planifiées. Votre administrateur peut advise vous que vous devez les tâches à effectuer avec Symantec AntiVirus.

Support technique

Dans le cadre de Symantec Security Response, le support technique Symantec Global met à jour dans toute la prise en charge de groupe centers. Le groupe de support technique rôle principal est en réponse à questions sur produit et d'auteur de contenu pour our la base de connaissances web-accessible. Le groupe de support technique collaboratively fonctionne avec les autres fonctionnel zones stockés dans Symantec pour answer votre questions dans un timely fashion. Par exemple, le groupe de support technique fonctionne avec d'autres groupes pour fournir les mises à jour de définitions de virus et des services des alertes pour propagations de virus et les alertes de sécurité.

Support technique Symantec offerings incluent:

- Un intervalle de prise en charge les options que vous donnent la flexibilité pour sélectionner la droite laps de service pour n'importe quel taille société
- Telephone et Web prennent en charge les composants qui fournissent rapid réponse et des informations up-to-the-minute
- Insurance la mise à niveau automatique de mise à niveau qui fournit le logiciel de protection
- Des mises à jour de contenu pour les définitions de virus et security-signatures que vous assurer la plus haut niveau de protection
- Global la prise en charge de Symantec Security Response, 24 heures un jour, 7 jours 'une semaine dans une série de langues pour ces enrolled dans le support Platinum programme.

Règles de cheval de Troie

Les chevaux de Troie sont Programmes malveillants disguised comme utile programmes. Quand vous installer et exécuter un cheval de Troie, elle apparaît pour effectuer une fonction, mais helpful endommage votre ordinateur système d'exploitation. Symantec Client Firewall cheval de Troie règles examiner les communications réseau de Symantec Client Firewall clients qui accèdent à Internet, recherchant signs de ces Programmes malveillants. Si l'un est détecté, la règle entre vision action contre ce type de menace. Cheval de Troie règles ont une priorité plus bas que générales et des règles de programme. Ils sont appliquées seulement après ces deux groupes de règles sont appliquées. Cheval de Troie règles par défaut toujours bloquer par opposition à générales et des règles de programme, qui peuvent permettre l'accès. Règles de cheval de Troie les configurations connues d'attaque par les correspondances de travail avec une liste des menaces connues ongoing contre les communications réseau. De temps en temps, inoffensif l'activité réseau peuvent déclencher une alerte, si la communication implique utilisant des ports spécifiques ou d'autres critères qui sont associés à un cheval de Troie connus. Si vous continually recevoir le même Alert, vous pouvez vous assurer que l'activité normal ou les communications sur votre réseau n'est pas générer l'alerte.

Supprimer les fichiers infectés par l'intermédiaire de la quarantaine

Si vous supprimez un fichier en quarantaine, Symantec AntiVirus de manière permanente supprime de votre ordinateur disque dur. Supprimer un fichier infecté réduit la menace qu'un virus peut se répandent en supprimant le fichier et virus de votre ordinateur. Supprimer le fichier infecté est utile pour les virus de fichier et les virus de macro. Puisque les virus peut dommages parties de un fichier, la suppression et remplaçant il avec un nettoyer un fichier de sauvegarde peut être meilleur que cleaning du fichier infecté. Vous pouvez effectuer cette action manuellement après un fichier infecté a été mis en quarantaine. Supprimer le fichier infecté en quarantaine serait un utile manière de supprimer un virus à partir d'un fichier qui a été disposable ne peut pas être nettoyé. Utilisez cette option seulement si vous avez nettoyer sauvegardes de que les fichiers que vous avez décidé à analyser. Vous devriez pas utiliser cette méthode comme Opération principale pour les fichiers qui sont analysés pendant Auto-Protect ou des analyses planifiées.

Pour activer et désactiver Auto-Protect

Si vous n'avez pas modifié les paramètres d'option par défaut, Auto-Protect Charge quand vous démarrez votre ordinateur efficace pour protéger contre les virus et les risques de sécurité. Il vérifie en cours d'exécution de programmes pour les virus et les risques de sécurité et Contrôles votre ordinateur pour tous les activités suspectes. Quand un virus, une activité suspecte (comportement pouvant être la présence d'un virus) ou un risque de sécurité est détecté, Auto-Protect vous alerte. Dans certains cas, Auto-Protect peut Avertir vous sur un rechercher les activités que vous savez qu'il n'est pas la présence d'un virus. Par exemple, cette avertissement peut se produire quand vous installez de nouveaux programmes. Si vous effectuez tels une activité et voulez éviter l'avertissement, vous pouvez désactiver Auto-Protect temporairement. Veillez à activer Auto-Protect quand vous avez terminé vos tâche pour s'assurer que votre ordinateur reste protégé. Votre administrateur peut verrouiller Auto-Protect de sorte que vous ne puissiez le désactiver, ou spécifier qu'elle peut être désactivé temporairement, mais reenables automatiquement après un délai spécifié.

À propos d'inclusions et des exclusions dans les analyses

A l'exclusion comprenant et les objets internes peuvent vous aider à équilibrer la quantité de protection requis avec le laps de ressources nécessaires pour fournir que la

protection. Si vous choisissez de ex analyser tous les types de fichier, vous pourriez vouloir exclure les dossiers contenant des fichiers de données qui ne sont pas est soumis aux virus. Autrement, vous pouvez analyser seulement les fichiers avec des extensions qui sont susceptibles un virus ou un risque de sécurité. Quand vous sélectionnez pour analyser seulement certaines extensions, vous excluez automatiquement tous les fichiers avec d'autres extensions de l'analyse. Ces choix diminuez la charge qui est associée à l'analyse des fichiers. Selon le type d'analyse et les objets internes de votre analyse, vous pouvez exclure par des fichiers, des dossiers des extensions de fichier ou types de fichier. Vous pouvez inclure seulement certains types de fichier ou extensions dans une analyse. Vous pouvez inclure et exclure des éléments des analyses que vous avez lancée de Symantec Client Security client ou le serveur de l'interface utilisateur ou depuis la console Symantec System Center.

E. Controlled Language Rule Set

1 General Style Rules

1.1.1 Keep the Subject and Verb Close to Each Other

Rule Name: verb close to subject

Keep the subject and verb close to each other at the beginning of a sentence.

1.1.2 Avoid Meaningless Openers

Rule Name: avoid meaningless openers

Sentences and clauses that begin with "there is" or "it is" are weak and wordy, because they include only one piece of information that does not reveal an interaction between two elements.

1.2 Do Not Compound Words

Rule Name: do not compound

Do not compound words such as past participles or adjectives.

1.3 Use a Hyphen to Indicate the First Part of a Compound

Rule Name: use hyphen in compound

Use a hyphen to indicate the first part of a compound that contains a past participle, such as "password-protected," "Web-based," "Windows-based," or "Windows-specific."

1.4 Do Not Omit Relative Pronouns Such as That and Which

Rule Name: use relative pronoun

Do not omit relative pronouns such as "that" and "which."

1.5 Use Complementizers

Rule Name: use complementizer

Do not omit "that" from a subordinate clause that contains an introductory verb. Examples of introductory verbs are "say," "tell," or "announce."

1.6 Do Not Omit Articles

Rule Name: use articles

Use articles for sentences in the following structures:
1) When nouns are defined by a restrictive relative clause;
2) When sentences begin with verbs that have no subjects.

1.8 Sentence too Long

Rule Name: sentence length

Restrict a sentence so that it expresses only one thought. Avoid sub-clauses when possible, except in obvious cases (such as conditional phrases introduced by "if"). Use no more than 24 words per sentence.

1.9 Avoid Using the Passive Voice

Rule Name: avoid passive

Use the active voice when possible. The active voice clarifies who or what is doing the action and is usually more direct and less wordy than the passive voice.
This version of the rules marks only the sentences that specify an agent.

1.10 Avoid Unnecessary Words

Rule Name: avoid unnecessary words

Avoid the following unnecessary words:

- "above", "absolute", "absolutely", "actually", "at this point", "basic", "below", "best, of, breed", "clearly", "dramatic", "extremely", "hugely", "just", "minimally", "nice", "obviously", "of course", "popular", "rarely", "realistically", "really", "simple", "simply", "state, of, the, art", "step, by, step", "strongly", "sufficiently", "unnecessarily", "virtually."

1.11 Place All Nontranslatable Text on Its Own Line

Rule Name: nontranslatable text on own line
Place all nontranslatable text on its own line.

1.12 Use the Serial Comma

Rule Name: use serial comma
Include the serial comma in a list of three or more items.

1.13 Avoid he, she, he/she, and s/he

Rule Name: avoid s he
Do not use he, she, he/she, and s/he.

1.14 Do Not Write the Full Name of Each Operating System

Rule Name: shorten OS reference
When you refer to multiple operating systems, do not write the full name of each operating system.

1.15.1 Do Not Use Future Tense

Rule Name: avoid future tense
Whenever possible, use the present tense rather than the future tense.
Occasionally, you may require a future tense because you are describing a future action.

1.15.2 Avoid Progressive Tense

Rule Name: avoid progressive tense
Avoid progressive tense: do not use a form of "be" followed by a participle.

1.16 Use Numerals for All Measurements Over 10

Rule Name: use numerals
Use numerals for all measurements over 10.

1.17 Repeat the Unit of Measure

Rule Name: repeat unit
For two or more quantities, repeat the unit of measure.

1.18 Use a Hyphen in a Unit

Rule Name: use hyphen in unit
When the measurement is used as an adjective, use a hyphen.

1.19 Use Number × Number

Rule Name: use number x number
Use number × number, not number by number.
The × should be delimited by one space on each side.

1.20 Avoid a Colon After a Drive Letter

Rule Name: avoid colon after drive
Do not use a colon after a drive letter, except when the drive letter is part of a path.

1.21 Do Not Use More Than Two Adverbs or Adjectives in a Series

Rule Name: avoid series of adjectives
Do not use more than two adverbs or adjectives in a series, with either a comma or "and" separating them.

1.22 Use a Noun at the Start of a Subordinate Clause

Rule Name: use noun in subordinate clause
Use a noun at the beginning of a subordinate clause.

1.23 Do Not Use 'this' or 'that' When They Are Not Followed by a Noun

Rule Name: use this that with noun
Use a noun after "this" and "that."

1.24 Punctuate Imperative Sentences in Bulleted Lists

Rule Name: punctuate imperative sentences in bulleted lists
Use end punctuation with imperative sentences in bulleted lists. Do not use end punctuation for incomplete bulleted phrases even if followed by a complete sentence.

1.25 Use Sentence-style Capitalization for Bulleted Lists

Rule Name: use capitalization in bulleted lists
Capitalize first word and proper nouns in bulleted lists.

1.26 Use a Colon at the End of a Sentence to introduce a Bulleted List

Rule Name: use colon before bulleted list
Use a colon at the end of a sentence to indicate that a bulleted list follows.

1.27 Write Positive Statements

Rule Name: write positive statements
Write positive statements. Do not use double negation.
This rule only applies in the context <warning>.

2 MT rules[1]

2.2 Repeat the Head Noun

Rule Name: repeat head noun
Repeat the head noun with conjoined articles or prepositions.

2.3 Do Not Use Slashes

Rule Name: avoid slashes
Do not use slashes to link common words.

2.4 Keep Both Parts of a Two-Part Verb Together

Rule Name: keep two verb parts together
Translation will be easier if you keep together both parts of a two-part verb.

2.5 Use "could" with "if"

Rule Name: could only with if
Use "could" only if the sentence contains a conditional clause that is introduced by "if."

2.6 Avoid "-ing" Words

---

[1] MT rules are only used for content that is to be sent to the MT system for translation.

Rule Name: avoid ing words

Whenever possible, avoid "-ing" words. They are highly ambiguous in English because they can be used as nouns, adjectives, and verbs. They are also difficult to translate.

2.7 Avoid Parenthetical Expressions in the Middle of a Sentence

Rule Name: avoid parenthetical expressions

Avoid parenthetical expressions in the middle of a sentence. Make the parenthetical expression a separate sentence if needs be.

3 Other Rules

3.1 Avoid Incomplete Segments

Rule Name: avoid incomplete segment

Make sure that every segment can stand alone.

F. Recall Test

[The instructions are shown below, the test contents are presented on the following pages: uncontrolled condition, then the controlled condition.]

Recall

Before we begin, we will explain what we would like you to do on the following pages:

You will be asked to answer three questions about each of the paragraphs you have just read. Each paragraph are treated in the same order as before and the title of the paragraphs is also given as a reminder.

1. The first question is a simple yes-no question about your overall comprehension of the paragraph. Please mark the appropriate answer.

2. The second question will ask you to fill in two/three blanks from a sentence you have seen in each of the paragraphs. The sentence is given in each case. Please fill in the missing words on the blanks or beside the sentence if you run out of space – each blank corresponds to ONE missing word and its length is no indication of the missing word's length.

3. The third, and final, question will ask you about more specific content of the paragraphs to test your in-depth comprehension. Please use the box provided for your answer and use space beside if necessary.

Don't worry! It's very straightforward and easy to complete - please take as much time as you need.

1. A propos de Symantec AntiVirus

1.1 After reading this paragraph, do you understand the options for installing Symantec AntiVirus? Yes No

1.2 Votre administrateur s'exécute analyses sur votre ordinateur et peut définir les _____ _____.'

1.3 What are the consequences of an administrator managing the installation?

2. Le support technique

2.1 After reading this paragraph, do you understand what tehnical support does in this context? Yes No

2.2 'Mise à niveau jour que delivers mise à niveau des logiciels de _____ _____.'

2.3 For which additional features are individuals who are enrolled on the Platinum Support Program eligible?

3. Règles de cheval de Troie

3.1 After reading this paragraph, do you understand the function of Trojan Horse rules? Yes No

3.2 Occasionnellement, inoffensif l'activité du réseau peuvent déclencher un _____ de _____ _____.'

3.3 What is the default function of a Trojan Horse rule?

4. Suppression de fichiers infecté s par des virus en quarantaine

4.1 After reading this paragraph, do you understand the purpose of Quarantine? Yes No

4.2 Utilisez cette option uniquement si vous avez Nettoyer sauvegardes de fichiers que vous decided _____ à _____.'

4.3 Why might it be better to replace a file with a clean back-up, rather than cleaning the infected file?

5. Activation et désactivation d 'Auto-Protect

5.1 After reading this paragraph, do you understand why it is necessary to enable or disable Auto-Protect? Yes No

5.2 Lorsqu'un virus, les activités suspectes (un événement qui ne peuvent être la présence d'un virus) ou un risque de sécurité est détecté, Auto-Protect ____ _____.'

5.3 What should you do after temporarily disabling Auto-Protect to perform a specific task?

6. A propos des inclusions et des exclusions lors des analyses

6.1 After reading this paragraph, do you understand why it is necessary to exclude items from scans? Yes No

6.2 Vous pouvez inclure seulement certaines types de fichier ou des extensions dans ___ _____.'

6.3 When you decide to scan one particular file extension, what is the consquence of this on other file extensions?

1. À propos de Symantec AntiVirus

1.1 After reading this paragraph, do you understand the options for installing Symantec AntiVirus? Yes No

1.2 Votre administrateur exécute des analyses sur votre ordinateur et peut configurer des _____ _____.'

1.3 What are the consequences of an administrator managing the installation?

2. Support technique

2.1 After reading this paragraph, do you understand what tehnical support does in this context? Yes No

2.2 Insurance la mise à niveau automatique de mise à niveau qui fournit le _____ de _____.'

2.3 For which additional features are individuals who are enrolled on the Platinum Support Program eligible?

3. Règles de cheval de Troie

3.1 After reading this paragraph, do you understand the function of Trojan Horse rules? Yes No

3.2 What is the default function of a Trojan Horse rule?

3.3 De temps en temps, inoffensif l'activité réseau peuvent déclencher ___ _____.'

4. Supprimer les fichiers infectés par l'intermédiaire de la quarantaine

4.1 After reading this paragraph, do you understand the purpose of Quarantine? Yes No

4.2 Utilisez cette option seulement si vous avez nettoyer sauvegardes de que les fichiers que vous avez _____ _ _____.'

4.3 Why might it be better to replace a file with a clean back-up, rather than cleaning the infected file?

5. Pour activer et désactiver Auto-Protect

5.1 After reading this paragraph, do you understand why it is necessary to enable or disable Auto-Protect? Yes No

5.2 Quand un virus, une activité suspecte (comportement pouvant être la présence d'un virus) ou un risque de sécurité est détecté, Auto-Protect ____ _____.'

5.3 What should you do after temporarily disabling Auto-Protect to perform a specific task?

6. À propos d’inclusions et des exclusions dans les analyses

6.1 After reading this paragraph, do you understand why it is necessary to exclude items from scans? Yes No

6.2 Vous pouvez inclure seulement certains types de fichier ou extensions dans ___ _____.'

6.3 When you decide to scan one particular file extension, what is the consquence of this on other file extensions?

G. Post-Task Evaluation

[The instructions are shown below, the test contents are presented on the following pages: uncontrolled condition, then the controlled condition.]

Evaluation

Before we begin, we will explain what we would like you to keep in mind when you are evaluating the texts that you have read earlier. We are asking you to evaluate the sentences of the paragraphs in terms of their readability and comprehensibility. The paragraphs are presented in the same order as before, but this time split into sentences. For the purposes of this test we define readability and comprehensibility as follows:

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)

Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

In other words, the higher the number on the scale, the better you could ***read*** and ***comprehend*** the sentence.

Simply mark the number relating to the sentence to judge how ***readable*** and ***comprehensible*** the sentence is. Here is an example:

1. Avant de passer à l'étape suivante, assurez-vous que le logiciel est mis à jour.

This sentence is readable.
1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

1 2 3 4 5

This sentence is comprehensible.
1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

1 2 3 4 5

Don't worry! The legend and explanations will appear on each page, so you won't need to memorise them. Take as much time as you need.

1	Readability	Comprehensibility
A propos de Symantec AntiVirus	1 2 3 4 5	1 2 3 4 5
Vous pouvez installer Symantec AntiVirus à la protection contre les virus et les risques de sécurité de l'une ou l'autre autonome ou une administrator-managed l'installation.	1 2 3 4 5	1 2 3 4 5
Autonome signifie que votre installation du logiciel Symantec AntiVirus n'est pas géré par un administrateur réseau.	1 2 3 4 5	1 2 3 4 5
Si vous gérer votre propre ordinateur, il doit être l'un des types suivants:	1 2 3 4 5	1 2 3 4 5
- Ordinateur autonome non connecté à un réseau, tel qu'un ordinateur d'accueil ou un portable autonome, avec l'installation de Symantec AntiVirus utilisant l'une ou l'autre de l'option par défaut paramètres ou de paramètres des options de administrator-preset	1 2 3 4 5	1 2 3 4 5
- Ordinateur distant qui sur votre réseau d'entreprise respectent qui doit exigences de sécurité avant la connexion	1 2 3 4 5	1 2 3 4 5
Le paramètre par défaut pour Symantec AntiVirus fournit une protection contre les virus et risques de sécurité pour votre ordinateur.	1 2 3 4 5	1 2 3 4 5
Toutefois, vous pouvez créés resuit jamais besoins de votre société, à optimize les performances système et de désactiver les options qui ne s'applique pas.	1 2 3 4 5	1 2 3 4 5
Si votre installation est géré par l'administrateur, certaines options peuvent être verrouillé ou non disponibles, ni risque de ne pas apparaître du tout, en fonction de votre politique de sécurité de son apportées par l'administrateur.	1 2 3 4 5	1 2 3 4 5
Votre administrateur s'exécute analyses sur votre ordinateur et peut définir les analyses planifiées.	1 2 3 4 5	1 2 3 4 5
Votre administrateur peut advise vous en tant que vous devez effectuer des tâches pour les éléments à l'aide de Symantec AntiVirus.	1 2 3 4 5	1 2 3 4 5

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)
Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

2	**Readability**	**Comprehensibility**
Le support technique	1 2 3 4 5	1 2 3 4 5
Dans le cadre de Symantec Security Response, Symantec Global Support technique gère la prise en charge du groupe de centers pendant toute la planète.	1 2 3 4 5	1 2 3 4 5
Le support technique de son groupe rôle essentiel de répondre aux questions spécifiques de produit fonctionnalité / fonctionner, l'installation et la configuration, ainsi que pour author pour de contenu de notre base de connaissances web-accessible.	1 2 3 4 5	1 2 3 4 5
Le support technique Fonctionnement du groupe collaboratively avec des zones vers d'autres functional dans Symantec answer votre questions dans une timely fashion.	1 2 3 4 5	1 2 3 4 5
Par exemple, le groupe de support technique fonctionnement engineering du produit, ainsi que Symantec Security Response pour fournir des services d'alerte et des mises à jour des définitions de virus épidémies de virus et les alertes de sécurité.	1 2 3 4 5	1 2 3 4 5
Le support technique de Symantec offerings sont les suivantes:	1 2 3 4 5	1 2 3 4 5
- Une plage de prendre en charge les options que vous donnez souplesse de droite pour sélectionner la quantité de service de tout société la taille.	1 2 3 4 5	1 2 3 4 5
- Telephone Web composants qui prennent en charge et fournit une réponse rapide et des informations up-to-the-minute	1 2 3 4 5	1 2 3 4 5
- Mise à niveau jour que delivers mise à niveau des logiciels de protection automatique	1 2 3 4 5	1 2 3 4 5
- Les mises à jour de contenu concernant les définitions de virus et les signatures de sécurité s'assurer que le plus haut niveau de protection	1 2 3 4 5	1 2 3 4 5
- Global Support de Symantec Security Response experts, disponible 24 heures un jour et 7 jours ouvrables dans un worldwide contiennent des langues pour ces enrolled pour prennent en charge les clients du programme	1 2 3 4 5	1 2 3 4 5

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)

Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

3	Readability	Comprehensibility
Règles de cheval de Troie	1 2 3 4 5	1 2 3 4 5
Les chevaux de Troie malveillants sont des programmes qui sont déguisées en utile pour les programmes.	1 2 3 4 5	1 2 3 4 5
Lorsque vous installez et exécutez un cheval de Troie, il apparaît pour être effectue une fonction utile alors mais il n'est jamais damaging votre ordinateur de système d'exploitation.	1 2 3 4 5	1 2 3 4 5
Symantec Client Firewall Règles de cheval de Troie examiner les communications réseau des clients Symantec Client Firewall qui accèdent à Internet, signe en recherchant de ces programmes dommageables.	1 2 3 4 5	1 2 3 4 5
Si elle est détectée, elle opération immédiate à ce type de menace.	1 2 3 4 5	1 2 3 4 5
Règles de cheval de Troie planifier la priorité inférieur à générales ou des règles de programme.	1 2 3 4 5	1 2 3 4 5
Elles sont appliques seulement une fois que ces deux groupes de règles sont appliquées.	1 2 3 4 5	1 2 3 4 5
Règles de cheval de Troie par défaut sont toujours le blocage des règles, dans contrast pour générales ou des règles de programme, qui peuvent autoriser l'accès.	1 2 3 4 5	1 2 3 4 5
Règles de cheval de Troie répar une correspondance de modèles d'attaque associé à la liste des menaces connues contre les communications réseau en permanence.	1 2 3 4 5	1 2 3 4 5
Occasionnellement, inoffensif l'activité du réseau peuvent déclencher un cheval de Troie Alert, si la communication implique d'utiliser des ports spécifiques ou d'autres connus critères associe à un cheval de Troie.	1 2 3 4 5	1 2 3 4 5
Si vous continually reçoivent les mêmes d'alerte de cheval de Troie, vous pouvez dé investigate dé pour imposer l'alerte n'a pas à génère par une activité normale ou les communications sur votre réseau.	1 2 3 4 5	1 2 3 4 5

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)
Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

4	**Readability**	**Comprehensibility**
Supprimer les fichiers infectés par l'intermédiaire de la quarantaine	1 2 3 4 5	1 2 3 4 5
Si vous supprimez un fichier en quarantaine, Symantec AntiVirus de manière permanente supprime de votre ordinateur disque dur.	1 2 3 4 5	1 2 3 4 5
Supprimer un fichier infecté réduit la menace qu'un virus peut se répandent en supprimant le fichier et virus de votre ordinateur.	1 2 3 4 5	1 2 3 4 5
Supprimer le fichier infecté est utile pour les virus de fichier et les virus de macro.	1 2 3 4 5	1 2 3 4 5
Puisque les virus peut dommages parties de un fichier, la suppression et remplaçant il avec un nettoyer un fichier de sauvegarde peut être meilleur que cleaning du fichier infecté.	1 2 3 4 5	1 2 3 4 5
Vous pouvez effectuer cette action manuellement après un fichier infecté a été mis en quarantaine.	1 2 3 4 5	1 2 3 4 5
Supprimer le fichier infecté en quarantaine serait un utile manière de supprimer un virus à partir d'un fichier qui a été disposable ne peut pas être nettoyé.	1 2 3 4 5	1 2 3 4 5
Utilisez cette option seulement si vous avez nettoyer sauvegardes de que les fichiers que vous avez décidé à analyser.	1 2 3 4 5	1 2 3 4 5
Vous devriez pas utiliser cette méthode comme Opération principale pour les fichiers qui sont analysés pendant Auto-Protect ou des analyses planifiées.	1 2 3 4 5	1 2 3 4 5

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)
Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree
In other words, the higher the number the greater amount of the sentence you could read and comprehend.

5	**Readability**	**Comprehensibility**
Activation et désactivation d 'Auto-Protect	1 2 3 4 5	1 2 3 4 5
Si vous n'avez pas modifié les paramètres d'option par défaut, Auto-Protect se charge au démarrage de l'ordinateur pour vous protéger contre les virus et les risques de sécurité. Vérifie les programmes pour rechercher les virus et les risques de sécurité car elles s'exécutent et contrôle de tout ordinateur activité peut indiquer la présence d'un virus ou d'un risque de sécurité.	1 2 3 4 5	1 2 3 4 5
Lorsqu'un virus, les activités suspectes (un événement qui ne peuvent être la présence d'un virus) ou un risque de sécurité est détecté, Auto-Protect vous alerte. Dans certains cas, Auto-Protect peut Avertir d'un concernant les activités suspectes qui vous savez qu'il n'est pas la présence d'un virus.	1 2 3 4 5	1 2 3 4 5
Par exemple, cela peut se produire si vous installez un nouveau programme sur l'ordinateur.	1 2 3 4 5	1 2 3 4 5
Si vous ne sera tel effectue une activité et créé viter l'avertissement, vous pouvez désactiver temporairement Auto-Protect.	1 2 3 4 5	1 2 3 4 5
Veillez à l'activez lorsque vous avez terminé votre tâche pour garantir que votre ordinateur reste protégés.	1 2 3 4 5	1 2 3 4 5
Votre administrateur peut verrouiller Auto-Protect pour que vous ne pouvez pas désactiver pour raison quelconque, ou spécifier que Auto-Protect pour le système de fichiers peuvent être désactivé, mais reenables temporairement automatiquement après une durée spécifiée.	1 2 3 4 5	1 2 3 4 5

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)
Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

6	**Readability**	**Comprehensibility**
A propos des inclusions et des exclusions lors des analyses	1 2 3 4 5	1 2 3 4 5
Des inclusions et des exclusions vous aider à balance la quantité de protection que votre réseau nécessite avec la durée et requis pour fournir des ressources cette protection.	1 2 3 4 5	1 2 3 4 5
Par exemple, si vous choisissez d'analyser tous les types de fichier, vous pouvez décider d'exclure certains dossiers contenant uniquement des fichiers de données qui ne peuvent pas être infectés.	1 2 3 4 5	1 2 3 4 5
Ou, il peut être utile de n'analyser que les fichiers portant des extensions qui sont susceptibles de contenir un virus ou un risque de sécurité.	1 2 3 4 5	1 2 3 4 5
Lorsque vous sélectionnez pour n'analyser que certaines extensions, vous excluez automatiquement tous les fichiers qui portent d'autres extensions de l'analyse.	1 2 3 4 5	1 2 3 4 5
Ces choix diminue le overhead associé à la recherche des fichiers.	1 2 3 4 5	1 2 3 4 5
Selon le type d'analyse et les objets de l'analyse, vous pouvez exclure par fichier, dossier ou type de fichier types de fichier.	1 2 3 4 5	1 2 3 4 5
Vous pouvez inclure seulement certaines types de fichier ou des extensions dans une analyse.	1 2 3 4 5	1 2 3 4 5
Vous pouvez inclure et exclure des à éléments des analyses lancées depuis Symantec Client Security de l'interface utilisateur client ou serveur ou depuis la console Symantec System Center.	1 2 3 4 5	1 2 3 4 5

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)
Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

1	Readability	Comprehensibility
À propos de Symantec AntiVirus	1 2 3 4 5	1 2 3 4 5
Vous pouvez installer Symantec AntiVirus™ de virus et de risque de sécurité en tant que la protection autonomes ou une installation gérée par l'administrateur.	1 2 3 4 5	1 2 3 4 5
Une installation autonome signifie qu'un administrateur réseau n'est pas gérer vos Symantec AntiVirus.	1 2 3 4 5	1 2 3 4 5
Si vous gérez vos propres ordinateur, il doit être un des types suivants:	1 2 3 4 5	1 2 3 4 5
- Un ordinateur autonome qui n'est pas connecté à un réseau par une installation de Symantec AntiVirus qui utilise les paramètres par défaut administrator-preset ou options	1 2 3 4 5	1 2 3 4 5
- Un ordinateur distant qui se connecte à votre réseau d'entreprise qui doivent répondre aux spécifications de sécurité avant connexion.	1 2 3 4 5	1 2 3 4 5
Les paramètres par défaut de Symantec AntiVirus assurent la protection de virus et de risque de sécurité pour votre ordinateur.	1 2 3 4 5	1 2 3 4 5
Cependant, vous pouvez régler adapter aux deux pour votre entreprise doit optimiser les performances du système et désactiver les options qui ne s'appliquent pas.	1 2 3 4 5	1 2 3 4 5
Si votre administrateur gère votre installation, quelques options peuvent être verrouillées ou indisponibles ou ne s'affiche pas à tout, selon la votre administrateur politique de sécurité.	1 2 3 4 5	1 2 3 4 5
Votre administrateur exécute des analyses sur votre ordinateur et peut configurer des analyses planifiées.	1 2 3 4 5	1 2 3 4 5
Votre administrateur peut advise vous que vous devez les tâches à effectuer avec Symantec AntiVirus.	1 2 3 4 5	1 2 3 4 5

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)
Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

2	**Readability**	**Comprehensibility**
Support technique	1 2 3 4 5	1 2 3 4 5
Dans le cadre de Symantec Security Response, le support technique Symantec Global met à jour dans toute la prise en charge de groupe centers.	1 2 3 4 5	1 2 3 4 5
Le groupe de support technique rôle principal est en réponse à questions sur produit et d'auteur de contenu pour our la base de connaissances web-accessible.	1 2 3 4 5	1 2 3 4 5
Le groupe de support technique collaboratively fonctionne avec les autres fonctionnel zones stockés dans Symantec pour answer votre questions dans un timely fashion.	1 2 3 4 5	1 2 3 4 5
Par exemple, le groupe de support technique fonctionne avec d'autres groupes pour fournir les mises à jour de définitions de virus et des services des alertes pour propagations de virus et les alertes de sécurité.	1 2 3 4 5	1 2 3 4 5
Support technique Symantec offerings incluent:	1 2 3 4 5	1 2 3 4 5
- Un intervalle de prise en charge les options que vous donnent la flexibilité pour sélectionner la droite laps de service pour n'importe quel taille société	1 2 3 4 5	1 2 3 4 5
- Telephone et Web prennent en charge les composants qui fournissent rapid réponse et des informations up-to-the-minute	1 2 3 4 5	1 2 3 4 5
- Insurance la mise à niveau automatique de mise à niveau qui fournit le logiciel de protection	1 2 3 4 5	1 2 3 4 5
- Des mises à jour de contenu pour les définitions de virus et security-signatures que vous assurer la plus haut niveau de protection	1 2 3 4 5	1 2 3 4 5
- Global la prise en charge de Symantec Security Response, 24 heures un jour, 7 jours 'une semaine dans une série de langues pour ces enrolled dans le support Platinum programme.	1 2 3 4 5	1 2 3 4 5

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)
Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

3	**Readability**	**Comprehensibility**
Règles de cheval de Troie	1 2 3 4 5	1 2 3 4 5
Les chevaux de Troie sont Programmes malveillants disguised comme utile programmes.	1 2 3 4 5	1 2 3 4 5
Quand vous installer et exécuter un cheval de Troie, elle apparaît pour effectuer une fonction, mais helpful endommage votre ordinateur système d'exploitation.	1 2 3 4 5	1 2 3 4 5
Symantec Client Firewall cheval de Troie règles examiner les communications réseau de Symantec Client Firewall clients qui accèdent à Internet, recherchant signs de ces Programmes malveillants.	1 2 3 4 5	1 2 3 4 5
Si l'un est détecté, la règle entre vision action contre ce type de menace.	1 2 3 4 5	1 2 3 4 5
Cheval de Troie règles ont une priorité plus bas que générales et des règles de programme.	1 2 3 4 5	1 2 3 4 5
Ils sont appliquées seulement après ces deux groupes de règles sont appliquées.	1 2 3 4 5	1 2 3 4 5
Cheval de Troie règles par défaut toujours bloquer par opposition à générales et des règles de programme, qui peuvent permettre l'accès.	1 2 3 4 5	1 2 3 4 5
Règles de cheval de Troie les configurations connues d'attaque par les correspondances de travail avec une liste des menaces connues ongoing contre les communications réseau.	1 2 3 4 5	1 2 3 4 5
De temps en temps, inoffensif l'activité réseau peuvent déclencher une alerte, si la communication implique utilisant des ports spécifiques ou d'autres critères qui sont associés à un cheval de Troie connus.	1 2 3 4 5	1 2 3 4 5
Si vous continually recevoir le même Alert, vous pouvez vous assurer que l'activité normal ou les communications sur votre réseau n'est pas générer l'alerte.	1 2 3 4 5	1 2 3 4 5

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)
Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

4	**Readability**	**Comprehensibility**
Supprimer les fichiers infectés par l'intermédiaire de la quarantaine	1 2 3 4 5	1 2 3 4 5
Si vous supprimez un fichier en quarantaine, Symantec AntiVirus de manière permanente supprime de votre ordinateur disque dur.	1 2 3 4 5	1 2 3 4 5
Supprimer un fichier infecté réduit la menace qu'un virus peut se répandent en supprimant le fichier et virus de votre ordinateur.	1 2 3 4 5	1 2 3 4 5
Supprimer le fichier infecté est utile pour les virus de fichier et les virus de macro.	1 2 3 4 5	1 2 3 4 5
Puisque les virus peut dommages parties de un fichier, la suppression et remplaçant il avec un nettoyer un fichier de sauvegarde peut être meilleur que cleaning du fichier infecté.	1 2 3 4 5	1 2 3 4 5
Vous pouvez effectuer cette action manuellement après un fichier infecté a été mis en quarantaine.	1 2 3 4 5	1 2 3 4 5
Supprimer le fichier infecté en quarantaine serait un utile manière de supprimer un virus à partir d'un fichier qui a été disposable ne peut pas être nettoyé.	1 2 3 4 5	1 2 3 4 5
Utilisez cette option seulement si vous avez nettoyer sauvegardes de que les fichiers que vous avez décidé à analyser.	1 2 3 4 5	1 2 3 4 5
Vous devriez pas utiliser cette méthode comme Opération principale pour les fichiers qui sont analysés pendant Auto-Protect ou des analyses planifiées.	1 2 3 4 5	1 2 3 4 5

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)
Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

5	**Readability**	**Comprehensibility**
Pour activer et désactiver Auto-Protect	1 2 3 4 5	1 2 3 4 5
Si vous n'avez pas modifié les paramètres d'option par défaut, Auto-Protect Charge quand vous démarrez votre ordinateur efficace pour protéger contre les virus et les risques de sécurité.	1 2 3 4 5	1 2 3 4 5
Il vérifie en cours d'exécution de programmes pour les virus et les risques de sécurité et Contrôles votre ordinateur pour tous les activités suspectes.	1 2 3 4 5	1 2 3 4 5
Quand un virus, une activité suspecte (comportement pouvant être la présence d'un virus) ou un risque de sécurité est détecté, Auto-Protect vous alerte.	1 2 3 4 5	1 2 3 4 5
Dans certains cas, Auto-Protect peut Avertir vous sur un rechercher les activités que vous savez qu'il n'est pas la présence d'un virus.	1 2 3 4 5	1 2 3 4 5
Par exemple, cette avertissement peut se produire quand vous installez de nouveaux programmes.	1 2 3 4 5	1 2 3 4 5
Si vous effectuez tels une activité et voulez éviter l'avertissement, vous pouvez désactiver Auto-Protect temporairement.	1 2 3 4 5	1 2 3 4 5
Veillez à activer Auto-Protect quand vous avez terminé vos tâche pour s'assurer que votre ordinateur reste protégé.	1 2 3 4 5	1 2 3 4 5
Votre administrateur peut verrouiller Auto-Protect de sorte que vous ne puissiez le désactiver, ou spécifier qu'elle peut être désactivé temporairement, mais reenables automatiquement après un délai spécifié.	1 2 3 4 5	1 2 3 4 5

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)
Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

6	**Readability**	**Comprehensibility**
À propos d'inclusions et des exclusions dans les analyses	1 2 3 4 5	1 2 3 4 5
A l'exclusion comprenant et les objets internes peuvent vous aider à équilibrer la quantité de protection requis avec le laps de ressources nécessaires pour fournir que la protection.	1 2 3 4 5	1 2 3 4 5
Si vous choisissez de ex analyser tous les types de fichier, vous pourriez vouloir exclure les dossiers contenant des fichiers de données qui ne sont pas est soumis aux virus.	1 2 3 4 5	1 2 3 4 5
Autrement, vous pouvez analyser seulement les fichiers avec des extensions qui sont susceptibles un virus ou un risque de sécurité.	1 2 3 4 5	1 2 3 4 5
Quand vous sélectionnez pour analyser seulement certaines extensions, vous excluez automatiquement tous les fichiers avec d'autres extensions de l'analyse.	1 2 3 4 5	1 2 3 4 5
Ces choix diminuez la charge qui est associée à l'analyse des fichiers. Selon le type d'analyse et les objets internes de votre analyse, vous pouvez exclure par des fichiers, des dossiers des extensions de fichier ou types de fichier.	1 2 3 4 5	1 2 3 4 5
Vous pouvez inclure seulement certains types de fichier ou extensions dans une analyse.	1 2 3 4 5	1 2 3 4 5
Vous pouvez inclure et exclure des éléments des analyses que vous avez lancée de Symantec Client Security client ou le serveur de l'interface utilisateur ou depuis la console Symantec System Center.	1 2 3 4 5	1 2 3 4 5

Readability: The extent to which the sentence is easy to read in terms of linguistic elements (grammar, structure, spelling – how it is being said)
Comprehensibility: The extent to which the content of the sentence is easy to understand (what is being said)

Legend: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

H. Source Text – Uncontrolled

About Symantec AntiVirus

You can install Symantec AntiVirus virus and security risk protection as either a stand-alone or an administrator-managed installation. A stand-alone installation means that your Symantec AntiVirus software is not managed by a network administrator. If you manage your own computer, it must be one of the following types: A stand-alone computer that is not connected to a network, such as a home computer or a laptop stand-alone, with a Symantec AntiVirus installation that uses either the default option settings or administrator-preset options settings; A remote computer that connects to your corporate network that must meet security requirements before connecting. The default settings for Symantec AntiVirus provide virus and security risk protection for your computer. However, you may want to adjust them to suit your company's needs, to optimize system performance, and to disable options that do not apply. If your installation is managed by your administrator, some options may be locked or unavailable, or may not appear at all, depending upon your administrator's security policy. Your administrator runs scans on your computer and can set up scheduled scans. Your administrator can advise you as to what tasks you should perform by using Symantec AntiVirus.

Technical Support

As part of Symantec Security Response, the Symantec global Technical Support group maintains support centers throughout the world. The Technical Support group's primary role is to respond to specific questions on product feature/function, installation, and configuration, as well as to author content for our Web-accessible Knowledge Base. The Technical Support group works collaboratively with the other functional areas within Symantec to answer your questions in a timely fashion. For example, the Technical Support group works with Product Engineering as well as Symantec Security Response to provide Alerting Services and virus definitions updates for virus outbreaks and security alerts. Symantec technical support offerings include: A range of support options that give you the flexibility to select the right amount of service for any size organization; Telephone and Web support components that provide rapid response and up-to-the-minute information; Upgrade insurance that delivers automatic software upgrade protection; Content Updates for virus definitions and security signatures that ensure the highest level of protection; Global support from Symantec Security Response experts, which is available 24 hours a day,

7 days a week worldwide in a variety of languages for those customers enrolled in the Platinum Support Program.

Trojan Horse Rules

Trojan horses are malicious programs that are disguised as useful programs. When you install and run a Trojan horse, it appears to be performing a helpful function, but it is actually damaging your computer's operating system. Symantec Client Firewall Trojan horse rules examine the network communications of Symantec Client Firewall clients that access the Internet, looking for signs of these malicious programs. If one is detected, the rule takes immediate action against this type of threat. Trojan horse rules have a lower priority than General or Program rules. They are applied only after those two groups of rules are applied. Default Trojan horse rules are always blocking rules, in contrast to General or Program rules, which may permit access. Trojan horse rules work by matching attack patterns associated with a list of known threats against ongoing network communications. Occasionally, harmless network activity can trigger a Trojan horse alert, if the communication involves using specific ports or other criteria associated with a known Trojan horse. If you continually receive the same Trojan horse alert, you may want to investigate further to make sure the alert is not being generated by normal activity or communications on your network.

Delete files that are infected by viruses in the Quarantine

If you delete a file in Quarantine, Symantec AntiVirus permanently deletes it from your computer's hard disk. Deleting a file that is infected by a virus reduces the threat that a virus might spread by removing the file (and thus the virus) from your computer. Deleting the infected file is useful for file viruses and macro viruses. Because viruses can damage parts of a file, deleting the infected file and replacing it with a clean backup file may be better than cleaning the infected file. You can perform this action manually after an infected file has been moved into the Quarantine. Deleting the infected file in the Quarantine would be a useful way to remove a virus from a disposable file that was unable to be cleaned. Use this option only if you have clean backups of files that you've decided to scan. You should not use this as a primary action for files that are scanned during Auto-Protect or scheduled scans.

Enabling and disabling Auto-Protect

If you have not changed the default option settings, Auto-Protect loads when you start your computer to guard against viruses and security risks. It checks programs for viruses and security risks as they run and monitors your computer for any activity that might indicate the presence of a virus or security risk. When a virus, virus-like activity (an event that could be the work of a virus), or security risk is detected, Auto-Protect alerts you. In some cases, Auto-Protect may warn you about a virus-like activity that you know is not the work of a virus. For example, this might occur when you are installing new computer programs. If you will be performing such an activity and want to avoid the warning, you can temporarily disable Auto-Protect. Be sure to enable Auto-Protect when you have completed your task to ensure that your computer remains protected. Your administrator might lock Auto-Protect so that you cannot disable it for any reason, or specify that File Auto-Protect can be disabled temporarily, but reenables automatically after a specified amount of time.

About inclusions and exclusions in scans

Inclusions and exclusions help you to balance the amount of protection that your network requires with the amount of time and resources that are required to provide that protection. For example, if you choose to scan all file types, you might want to exclude certain folders that contain only data files that are not subject to viruses. Or, you might want to scan only the files with extensions that are likely to contain a virus or other risk. When you select to scan only certain extensions, you automatically exclude all files with other extensions from the scan. These choices decrease the overhead that is associated with scanning files. Depending on the type of scan and the objects of your scan, you can exclude by files, folders, file extensions, or file types. You can include only certain file types or extensions in a scan. You can include and exclude items from scans that you initiate from the Symantec Client Security client or server user interface, or from the Symantec System Center console.

I. Source Text – Controlled

About Symantec AntiVirus

You can install Symantec AntiVirus virus and security risk protection as either a stand-alone or an administrator-managed installation. A stand-alone installation means that a network administrator does not manage your Symantec AntiVirus. If you manage your own computer, it must be one of the following types: A stand-alone computer that is not connected to a network with a Symantec AntiVirus installation that uses either the default or administrator-preset options settings; A remote computer that connects to your corporate network that must meet security requirements before it connects; The default settings for Symantec AntiVirus provide virus and security risk protection for your computer. However, you may want to adjust them to suit your company's needs, to optimize system performance, and to disable the options that do not apply. If your administrator manages your installation, some options may be locked or unavailable, or may not appear at all, depending upon your administrator's security policy. Your administrator runs scans on your computer and can set up scheduled scans. Your administrator can advise you as to what tasks you should perform by using Symantec AntiVirus.

Technical Support

As part of Symantec Security Response, the Symantec global Technical Support group maintains support centers throughout the world. The Technical Support group's primary role is to respond to questions on product and to author content for our Web-accessible Knowledge Base. The Technical Support group works collaboratively with the other functional areas within Symantec to answer your questions in a timely fashion. For example, the Technical Support group works with other groups to provide Alerting Services and virus definitions updates for virus outbreaks and security alerts. Symantec technical support offerings include: A range of support options that give you the flexibility to select the right amount of service for any size organization; Telephone and Web support the components that provide rapid response and up-to-the-minute information; Upgrade the insurance that delivers automatic software upgrade protection; Content Updates for virus definitions and security-signatures that ensure the highest level of protection. Global support from Symantec Security Response, 24 hours a day, 7 days a week in a variety of languages for those enrolled in the Platinum Support Program.

Trojan Horse Rules

Trojan horses are malicious programs disguised as useful programs. When you install and run a Trojan horse, it appears to perform a helpful function, but damages your computer's operating system. Symantec Client Firewall Trojan horse rules examine the network communications of Symantec Client Firewall clients that access the Internet, looking for signs of these malicious programs. If one is detected, the rule takes immediate action against this type of threat. Trojan horse rules have a lower priority than General or Program rules. They are applied only after those two groups of rules are applied. Default Trojan horse rules always block, in contrast to General or Program rules, which may permit access. Trojan horse rules work by matching the attack patterns with a list of known threats against ongoing network communications. Occasionally, harmless network activity can trigger an alert, if the communication involves using specific ports or other criteria which are associated with a known Trojan horse. If you continually receive the same alert, you may want to ensure that normal activity or communications on your network does not generate the alert.

Delete the infected files via Quarantine

If you delete a file in Quarantine, Symantec AntiVirus permanently deletes it from your computer's hard disk. Deleting an infected file reduces the threat that a virus might spread by removing the file and virus from your computer. Deleting the infected file is useful for file viruses and macro viruses. Because viruses can damage parts of a file, deleting and replacing it with a clean backup file may be better than cleaning the infected file. You can perform this action manually after an infected file has been moved into the Quarantine. Deleting the infected file in the Quarantine would be a useful way to remove a virus from a disposable file that was unable to be cleaned. Use this option only if you have clean backups of that files that you've decided to scan. You should not use this method as a primary action for the files that are scanned during Auto-Protect or scheduled scans.

To enable and disable Auto-Protect

If you have not changed the default option settings, Auto-Protect loads when you start your computer to guard against viruses and security risks. It checks running programs for viruses and security risks and monitors your computer for any suspicious activity. When a virus, virus-like activity (an event that may be the work of a virus), or security risk is detected, Auto-Protect alerts you. In some cases, Auto-Protect may warn you about a virus-like activity that you know is not the work of a virus. For example, this warning might occur when you

install new computer programs. If you perform such an activity and want to avoid the warning, you can temporarily disable Auto-Protect. Be sure to enable Auto-Protect when you have completed your task to ensure that your computer remains protected. Your administrator might lock Auto-Protect so that you cannot disable it, or specify that it can be disabled temporarily, but reenables automatically after a specified time.

About inclusions and exclusions in scans

Including and excluding objects can help you to balance the amount of required protection with the amount of resources necessary to provide that protection. E.g. if you choose to scan all file types, you might want to exclude folders containing the data files that are not subject to viruses. Otherwise, you might want to scan only the files with the extensions that are likely to contain a virus or other risk. When you select to scan only certain extensions, you automatically exclude all files with other extensions from the scan. These choices decrease the overhead that is associated with scanning files. Depending on the type of scan and the objects of your scan, you can exclude by files, folders, file extensions, or file types. You can include only certain file types or extensions in a scan. You can include and exclude items from the scans that you initiate from the Symantec Client Security client or server user interface, or from the Symantec System Center console.