MEASURING DIFFICULTY IN ENGLISH-CHINESE TRANSLATION: TOWARDS
A GENERAL MODEL OF TRANSLATION DIFFICULTY

A dissertation submitted
to Kent State University in partial
fulfillment of the requirements for the
degree of Doctor of Philosophy

by
Sanjun Sun
August 2012
Dissertation written by

Sanjun Sun

M.A., Beihang University, China, 2005

B.A., Ocean University of China, China, 1999

Approved by

___________________________________, Chair, Doctoral Dissertation Committee
Gregory M. Shreve

___________________________________, Member, Doctoral Dissertation Committee
Carol Maier

___________________________________, Member, Doctoral Dissertation Committee
Erik B. Angelone

___________________________________, Member, Doctoral Dissertation Committee
Jocelyn R. Folk

___________________________________, Graduate Faculty Representative
Andrew Barnes

Accepted by

___________________________________, Chair, Department of Modern & Classical Language Studies
Jennifer Larson

___________________________________, Dean, College of Arts and Sciences
John R.D. Stalvey
TABLE OF CONTENTS

LIST OF FIGURES ........................................................................................................ VI
LIST OF TABLES ....................................................................................................... VIII
ACKNOWLEDGEMENTS ........................................................................................... IX
CHAPTER 1 INTRODUCTION ..................................................................................... 1

CHAPTER 2 LITERATURE REVIEW ......................................................................... 5

2.1 Difficulty, mental workload and cognitive load...................................................... 5

2.2 Sources of translation difficulty .............................................................................. 10

2.2.1 Translation factors........................................................................................... 10

2.2.1.1 Text difficulty ......................................................................................... 11

2.2.1.2 Translation-specific difficulty ................................................................ 16

2.2.2 Translator factors........................................................................................... 20

2.3 Measuring translation difficulty ........................................................................... 29

2.3.1 Measuring text readability................................................................................. 29

2.3.2 Recognizing translation problems and difficulty ........................................... 36

2.3.2.1 Grading translations ............................................................................... 38

2.3.2.2 Analyzing verbal protocols ...................................................................... 43

2.3.2.3 Recording and analyzing translation behavior ....................................... 44

2.3.3 Measuring mental workload ........................................................................... 46

2.3.3.1 Subjective measures ............................................................................... 46

2.3.3.2 Performance measures .......................................................................... 50
2.3.3.3 Physiological measures ................................................................. 52

CHAPTER 3 METHODS .............................................................................. 55

3.1 Participants ......................................................................................... 55

3.2 Materials ............................................................................................ 56

3.2.1 Test passages ............................................................................... 57

3.2.2 Translation ratings ........................................................................ 59

3.3 Design and procedure ....................................................................... 61

3.3.1 Experimental design ................................................................. 61

3.3.2 Test procedure ............................................................................... 62

3.3.3 Pilot study ...................................................................................... 63

3.3.4 Grading the translations ............................................................ 65

3.3.5 Calculating interrater reliability ............................................... 69

3.3.6 Checking for normality and outliers .............................................. 70

CHAPTER 4 RESULTS ............................................................................. 74

4.1 Checking NASA-TLX’s reliability .................................................... 74

4.2 Translation difficulty level and translation quality score ................ 79

4.3 Translation difficulty level and time on task .................................. 83

4.4 Translation difficulty level and text readability ............................. 84

4.5 Predicting the level of translation difficulty of a text ..................... 88

4.6 Sources of translation difficulty in terms of translation errors ...... 92

CHAPTER 5 DISCUSSION ......................................................................... 99
5.1 Discussion of Findings ........................................................................................................ 99
  5.1.1 Readability formulas and translation difficulty ......................................................... 100
  5.1.2 Good translators and poor translators ................................................................. 102
  5.1.3 Workload and performance .............................................................................. 104
  5.1.4 Translation competence .................................................................................... 107

5.2 Limitations of this study and suggestions for further study ......................... 108

APPENDIX A TEST PASSAGES ..................................................................................... 112
APPENDIX B THE POST-TRANSLATION QUESTIONNAIRE ..................................... 118
APPENDIX C THE ADAPTED NASA TASK LOAD INDEX FOR MEASURING
  TRANSLATION DIFFICULTY ................................................................................... 119
APPENDIX D DATA FLOWCHART FOR ERROR POINT DECISIONS (VERSION
  2009) .......................................................................................................................... 120
APPENDIX E HOLISTIC TRANSLATION SCORING CRITERIA .............................. 121
APPENDIX F STATGRAPHICS OUTPUT FOR THE STEPWISE MULTIPLE
  REGRESSION ANALYSIS ......................................................................................... 122
REFERENCES .............................................................................................................. 125
LIST OF FIGURES

Fig. 2.1 A NASA-TLX subscale........................................................................................................ 48
Fig. 2.2 Hypothetical relationship between workload and performance ......................... 51
Fig. 3.1 Readability Studio .......................................................................................................... 59
Fig. 3.2 The pre-translation rating ............................................................................................ 60
Fig. 3.3 Original NASA-TLX subscale .................................................................................... 64
Fig. 3.4 Adapted NASA-TLX subscale .................................................................................... 64
Fig. 3.5 Original Performance subscale in NASA-TLX .......................................................... 64
Fig. 3.6 Adapted Performance subscale in NASA-TLX .......................................................... 65
Fig. 3.7 Error category list used in grading ............................................................................ 67
Fig. 3.8 The interface of TransGrading .................................................................................... 68
Fig. 4.1 High scorers vs. low scorers in translation difficulty score ........................................ 81
Fig. 4.2 Correlation between translation quality score and translation difficulty score ... 82
Fig. 4.3 Correlation between translation difficulty level and time on task ......................... 84
Fig. 4.4 Relationship between passage category and translation difficulty score .......... 87
Fig. 4.5 Translation difficulty score as a function of the Flesch score .............................. 88
Fig. 4.6 Relationship between self-predicted translation difficulty level and the actual translation difficulty level ...................................................................................... 92
Fig. 4.7 Translation error categories and points deducted .................................................... 95
Fig. 4.8 Frequency of translation errors and points deducted .............................................. 96
Fig. 4.9 Source text comprehension vs. reverbalization in terms of difficulty................. 98
# LIST OF TABLES

Table 2.1 SAE J2450 Translation Quality Metric ................................................................. 40
Table 3.1 Participants in this study ..................................................................................... 56
Table 3.2 Test passages and their Flesch Reading Ease scores ......................................... 58
Table 3.3 Part of the test passage order arrangement ....................................................... 61
Table 3.4 Passage numbers ............................................................................................... 61
Table 3.5 Descriptive statistics for the 7 variables ........................................................... 71
Table 4.1 Descriptive statistics for the 15 passages .......................................................... 76
Table 4.2 Item-total statistics for NASA-TLX’s internal consistency ............................. 77
Table 4.3 Undergraduate versus graduate students in their mean translation quality score ................................................................................................................................... 79
Table 4.4 Time on translation and Translation quality score ............................................ 83
Table 4.5 Readability scores of the 15 passages ............................................................... 85
Table 4.6 Descriptive statistics for the three passage categories ...................................... 86
Table 4.7 Mean translation difficulty scores of the 15 passages ..................................... 89
Table 4.8 Translation error categories and corresponding translation subcompetences .. 93
Table 4.9 Translation subcompetence categories and error points deducted............... 96
ACKNOWLEDGEMENTS

In completion of this doctoral dissertation, I would like to express my thanks and appreciation to many people. First, I would like to thank my advisor, Dr. Gregory M. Shreve, who provided me with valued guidance and support throughout the completion of this dissertation. I would also like to thank Drs. Carol Maier, Erik B. Angelone, and Jocelyn R. Folk for their willingness to serve on my dissertation committee and for their valuable and helpful contributions to this project. I wish to thank Dr. Isabel Lacruz, who helped me design the experiment, Drs. Wen Jun, Wang Keyou, and Wang Ying for helping me collect data. In addition, I would like to thank Drs. Françoise Massardier-Kenney, Brian Baer, Sue Ellen Wright, Geoffrey Koby, Keiran Dunne, Judy Wakabayashi, and Richard Kelly Washbourne for their assistance and encouragement.

I would like to thank all of my fellow doctoral students and friends for their friendship and encouragement. I am especially grateful to Luan Lingyan, Yu Zhongbo, and Ma Ji for providing unwavering support. Without their help, my life in the past four years would not have been so easy.

Finally, I would like to thank my family whose continued and enduring support throughout my life and especially in the pursuit of my educational goals has enhanced my ability to achieve these goals. I would also like to thank my ThinkPads (IBM/Lenovo) and Google, which were with me most of my waking hours when I was working on the dissertation. For me, they are part of my family.
CHAPTER 1

Introduction

The purpose of this study is to determine the factors causing translation difficulty and to explore ways to measure the translation difficulty level of source texts (English in this case) objectively and effectively. Eventually, an application for measuring translation difficulty can be developed based on discoveries to be made in this study and further studies.

Knowing the difficulty level of a translation assignment is important in translation pedagogy, accreditation and research as well as for the language industry. Traditionally, people rely on their holistic intuition to gain an idea of the level of a text’s translation difficulty. Although experts’ intuition (especially an expert panel’s judgment) is “reasonably reliable” (Campbell & Hale, 1999), we still need instruments or established procedures to make the evaluation process more effective and the results more objective.

In the field of pedagogy, Gickling & Rosenfield (1995) recommend when discussing best practices in curriculum-based assessment that “students’ accuracy [in comprehension] during reading instruction should be between 93 and 97%,” and “[w]hen task difficulty level is appropriate, other problems are forestalled” (from Daly, Chafouleas, & Skinner, 2005, p. 14). The implication for translation training is that the difficulty level of source texts should be appropriate for students. Hence, there is a need for properly leveled passages for translation exercises in translation pedagogy.
Assessing the translation difficulty level of testing materials is a must for translation accreditation bodies. Unfortunately, because of lack of research in this field, their assessments are not always as effective as intended. For instance, this author found, when studying American Translators Association (ATA)’s 2006 English-Chinese examinations, that, among the three test passages A (general), B (in the domain of medicine), and C (in the domain of business), passage C was the most difficult one to comprehend according to the readability test results, and the passing rate for passage C was the lowest among the passages. As ATA intended the passages to be equally difficult, this finding makes one doubt that their workgroup method for evaluating the translation difficulty level of those texts in 2006 was effective.

In translation research, translation difficulty deserves more attention. For instance, in process-oriented research, researchers have no standards to refer to when they choose test passages, and the texts used are diverse in terms of text type, length and, possibly, difficulty (see Krings, 2001, p. 74). This makes it hard for one to evaluate the comparability of experimental results between these studies. For instance, the use of different translation strategies in terms of type and frequency might vary depending on the translation difficulty level of the texts. Dragsted (2004) found in her empirical study that professional translators would adopt a more novice-like behavior during translation of a difficult text than during the translation of an easy text. Thus, translation difficulty may be an important variable in translation process research.

Translation difficulty research is also useful for the language industry. For instance, language service providers and translators can base their rates on the translation
difficulty level of clients’ materials. An established objective instrument will help two parties reach a consensus more easily.

To date, there have been few studies on translation difficulty. Campbell and Hale (e.g., 1999, 2002) and Jensen (2009) are among the few researchers who have done some exploratory empirical work in this direction. Hale & Campbell (2003) argue that it is possible to compile an inventory of universal sources of translation difficulty in source texts as well as lists for specific language pairs. By contrast, in the field of reading, researchers have been working on readability formulas for measuring text readability since the 1920s (for an overview, see Zamanian & Heydari, 2012).

To measure translation difficulty, we can measure language features (such as lexical complexity, syntactic complexity, and cohesion) using text analysis tools, and test translators using texts whose language features have been described. Then, multiple regression, a statistical technique for estimating relationships between variables, can be used to see which language features are correlated with translators’ objective performance (including time-on-task and translation quality score) and their subjective evaluations of translation difficulty. For assessing translation difficulty subjectively, NASA Task Load Index (TLX), a multidimensional scale for measuring subjective workload, can be used, but its reliability needs to be tested. Readability formulas measure text complexity, and there is a possibility that they can be used to measure translation difficulty (Jensen, 2009).

This study focused on the following research questions:

1) Is NASA-TLX reliable in measuring translation difficulty?
2) Can translation quality score alone be used to measure or represent translation difficulty?

3) Can time on task be used to measure or represent translation difficulty?

4) Can Flesch Reading Ease formula (or readability formulas in general) be used to predict a text's level of translation difficulty?

5) How can we know a text's level of translation difficulty without having the text translated first?

6) What are the sources of translation difficulty in terms of translation errors?

Chapter 2 of this dissertation describes the relevant literature pertaining to translation difficulty, from which the research questions are derived, and lays a theoretical foundation for this line of inquiry. Chapter 3 describes the study methods in detail, including participants, test materials, testing procedure, and grading procedures used in the study. Chapter 4 offers the results of the analyses related to each of the six research questions listed above. Chapter 5 discusses the study's findings, limitations, and possible directions for future research.
CHAPTER 2

Literature Review

Although the expression “translation difficulty” has been mentioned frequently in translation literature, few researchers have attempted to measure translation difficulty. This is a relatively new research field. It has two basic research questions with respect to difficulty: what to measure and how to measure it. Breaking up the research questions will enable us to draw on theories and findings in many other research fields as well as in translation studies. This chapter will explore these two questions and try to lay a theoretical foundation for this line of research. To begin, we need to know what difficulty is.

2.1 Difficulty, mental workload and cognitive load

From the cognitive perspective, difficulty refers to the amount of cognitive effort required to solve a problem. Difficult tasks need more processing effort. Difficulty is similar to complexity, but the two are not the same. According to Dahl (2004, p. 39), difficulty is a notion that primarily applies to tasks, and is always relative to person, hence the term “task difficulty;” complexity as an information-theoretic notion is more objective in the sense of being independent of the use. The construct of task difficulty in this sense has been adopted in such fields as job performance (e.g., Madden, 1962) and second language acquisition (e.g., Robinson, 2011).
Mental workload and cognitive load are similar to difficulty, and yet they have different theoretical backgrounds. The two will be discussed in the following paragraphs.

Mental workload has been an important concept in the field of human factors and ergonomics (see, e.g., Moray, 1979), which is concerned with the design and evaluation of tasks, products, and systems in order to make them compatible with the needs, abilities, and limitations of people (Karwowski, 2006, p. i). To date, there is still no universally accepted definition of mental workload. Hart & Staveland (1988, p. 140) define it as “a hypothetical construct that represents the cost [i.e., the amount of cognitive resources] incurred by a human operator to achieve a particular level of performance.” From a different theoretical perspective and on the basis of a comprehensive literature review, Young & Stanton (2006, p. 818) propose an operational definition of mental workload: “The mental workload of a task represents the level of attentional resources required to meet both objective and subjective performance criteria, which may be mediated by task demands, external support and past experience.” Gopher & Donchin (1986, p. 11) note that mental workload is an attribute of the person-task loop, and the effects of workload on performance can only be examined in relation to a model of human information processing.

It is generally accepted that mental workload is a multidimensional construct. A cohesive and multidimensional model incorporating factors affecting mental workload was proposed by Meshkati (1988). This model consists of two major sections—Causal Factors and Effect Factors—each with two primary component groups. The Causal Factors are 1) Task and Environmental Variables, which include task criticality, physical...
and psychological environmental factors, intrinsic task-related variables (including amount of information); time pressure, task structure and its rigidity (e.g., decoding requirements, decision making vs. problem solving), task novelty (to the operator); task rate/frequency; equipment used; and type of reward system, and 2) Operators’ Characteristics and Moderating Variables, which include cognitive capabilities of the individual (e.g., intellect); motivational states and personal utility system (e.g., goals, feedback orientation, attitude toward task and the utility of the task and its outcomes); past experience and training, etc. The Effect Factors include 1) task difficulty, response and performance, and 2) mental workload measures. These factors interact with each other.

Another model influential in the field of mental workload is the multiple resource model, proposed by Wickens (e.g., 1980, 1992, 2008). As this model was developed based on theories of attention, it is also called the multiple resource model of attention. Multiple attentional resources were identified along four fundamental dimensions: 1) stages of processing (distinguishing between perception, central processing, and responding stages); 2) codes of processing (distinguishing between spatial and verbal activity); 3) modalities (distinguishing between visual and auditory on the perception end of processing, and manual and vocal on the response end); and 4) visual channels (distinguishing between focal and ambient vision). In this four-dimensional model, each cell has resources dedicated to it. If two tasks use completely different cells, for instance, with one task requiring visual and manual processing, and with the second task requiring auditory and verbal processing, the two tasks will not interfere with each other's
performance; as resource overlap increases, interference will generally increase (Boles, 2006). In other words, “the higher the similarity in the resource demands among the task components, the more severe the resource competition and the lower the level of performance” (Tsang, 2006, p. 809).

The term “cognitive load” has been used in psychology since as early as the 1960s (e.g., Bradshaw, 1968). It can be generally defined as “the demand for working memory resources required for achieving goals of specific cognitive activities in certain situations” (Kalyuga, 2009, p. 35), although scholars in different fields have different ideas about this term. For instance, researchers in the field of human factors and ergonomics (e.g., Sammer, 2006) take cognitive load (as well as informational load, attentional load, and emotional load) as a subconcept of workload.

Cognitive load theory (CLT) was first advanced in 1988 by Sweller. It has attracted a group of researchers in educational psychology and applied learning sciences, and has been expanded significantly ever since (see, e.g., Sweller, Ayres, & Kalyuga, 2011). This line of research originated from the studies of human learning during performance of problem-solving tasks (e.g., Sweller, Mawer, & Howe, 1982). Its objective is to predict learning outcomes by taking into consideration the capabilities and limitations of human cognition (Plass, Moreno, & Brünken, 2010, p. 1), and it is based on the following assumption: a limited working memory makes it difficult to assimilate multiple elements of information simultaneously, but under conditions where multiple elements of information interact, they must be assimilated simultaneously; as a result, a heavy cognitive load is imposed (Sweller & Chandler, 1994).
According to CLT, instructional materials should be created for an optimal cognitive load, as cognitive overload impedes learning while cognitive underload does not generate interest (Lohr & Gall, 2008). An optimal cognitive load is a function of the individual learner’s expertise level (especially prior knowledge and effective working memory size), the instructional material’s complexity and the specific task. The magnitude of cognitive load is determined by the degree of interactivity between the three factors.

Compared with the mental workload construct, which covers dozens of factors (as mentioned above), the cognitive load construct in CLT has mainly focused on how the objective characteristics of the task affect cognitive load and learning, and students’ prior knowledge is the only individual characteristic explicitly included in its theoretical framework (Moreno & Park, 2010). From this perspective, the cognitive load construct is narrower than the mental workload construct.

Despite their differences in research scope and purpose, task difficulty, mental workload, and cognitive load are more similar than different. They all take into consideration the information processing demands that a certain task imposes on an individual and the limits of working memory. Compared with the other two constructs, mental workload is far more established in terms of impact and number of researchers in the field. Thus, mental workload theories can provide the major theoretical basis for translation difficulty research, supported by the other two constructs and related theories. As the term difficulty is more commonly used than “mental workload” or “cognitive load” in translation studies, it can serve as a cover term for this concept.
2.2 Sources of translation difficulty

Based on the discussion in the preceding section (especially the mental workload model proposed by Meshkati), sources of translation difficulty can be divided into two groups: task (i.e., translation) factors and translator factors. They are the factors that need to be measured.

2.2.1 Translation factors

The translation process is usually represented in a two-phase or three-phase model (Nord, 2005). In the two-phase model, the translation process consists of reading (also called “comprehension”, “decoding”, “analysis”) and reverbalization (also called “writing”, “encoding”, “recoding”, “rendering”, “reformulation”, “retextualization”, “synthesis”) (e.g., Neubert & Shreve, 1992; Wilss, 1982). The translator reads the source text in the first phase, and reverbalizes the meaning of the source text in the target language in the second phase. In the three-phase model, the translation process is divided into three stages: reading, transfer, and reverbalization. When translating, the translator first analyzes the message of the source language into its simplest forms, transfers it at this level, and then re-express it in the receptor language (Nida 1975, p. 79-80).

The following sections will discuss factors that may cause difficulty in translation. For the sake of discussion, the two-phase model will be adopted here. Reading comprehension is a broad research field (for an overview, see, e.g., Kamil, Pearson, Moje, & Afflerbach, 2011), and the domain that is particularly pertinent to translation difficulty research is text difficulty. In order to make the present discussion more accessible, factors that may cause difficulty in the second phase will be grouped under
the heading “translation-specific difficulty.” Between these two groups of factors there are unavoidable overlaps.

2.2.1.1 Text difficulty

Text difficulty and readability are a very important topic in reading research. Like translation researchers, reading researchers are also concerned with assessments, standards, and curriculum.

The RAND Reading Study Group (2002), a 14-member panel funded by the United States Department of Education's Office of Educational Research and Improvement, propose the following categories and dimensions that vary among texts and create varying challenges for readers (p. 25):

- Discourse genre, such as narration, description, exposition, and persuasion.
- Discourse structure, including rhetorical composition and coherence.
- Media forms, such as textbooks, multimedia, advertisements, hypertext, and the Internet.
- Sentence difficulty, including vocabulary, syntax, and the propositional text base (the explicit meaning of the text’s content drawn from propositions in the text, i.e., statements or idea units, but without more-subtle details about verb tense and deictic references [here, there, now, then, this, that]).
- Content, including different types of mental models, cultures, and socioeconomic strata; age-appropriate selection of subject matter; and the practices that are prominent in the culture.
- Texts with varying degrees of engagement for particular classes of readers.
Most of these categories are also mentioned by Gray & Leary in their book *What Makes a Book Readable* (1935). Based on a survey among about 100 librarians, publishers, and teachers and directors of adult classes, Gray & Leary compiled a list of 289 factors that may be related to a book’s readability, and classified them into four major categories (p. 295-300):

1) format or mechanical features, including size of book, number of pages, quality of paper, kind of type and printing, length of line, margin, general appearance of page, binding, and illustrations.

2) general features of organization, including title of book, chapter divisions, paragraph divisions, and reference guides.

3) style of expression and presentation, including vocabulary, sentences, paragraphs, chapters, attitude of author, method of presentation, style of presentation, and stylistic devices.

4) content, including theme, nature of subject matter, and unity of content

These 100 participants rated factors of content of greatest importance; factors of style, second; factors of format, third; and general features of organization of least importance. They believed that a readable book, first of all, must contain content that attracts and holds the reader's interest. A further survey among 170 readers showed that factors pertaining to style were most important for a book’s readability; factors of content, format, and factors of organization ranked second, third, and fourth, respectively. In other
words, “style of expression and presentation” (mainly lexical and syntactic complexity), and content are very important for the reader.

**Lexical and syntactic complexity**

Gray & Leary tested the reading ability of 1,690 adults, and found through statistical analysis that the best estimate of a text’s difficulty involved the use of eight elements: number of different hard words, number of easy words, percentage of monosyllables, number of personal pronouns, average sentence length in words, percentage of different words, number of prepositional phrases, and percentage of simple sentences (p. 16). These are all structural elements in the style group, as they “lend themselves most readily to quantitative enumeration and statistical treatment” (p. 7).

**Content and subject matter**

In terms of content and subject matter, it is commonly believed that abstract texts (e.g., philosophical texts) will be harder to understand than concrete and imaginable texts describing real objects, events or activities (e.g., stories), and texts on everyday topics are likely to be easier to process than those that are not (Alderson, 2000, p. 62). This can be explained by schema theory, which has been an important theory in reading comprehension since the 1980s.

According to Bartlett in his classic book *Remembering* (1932), the term "schema" refers to “an active organization of past reactions, or of past experiences, which must always be supposed to be operating in any well-adapted organic response” (p. 201). From this theory's perspective, the reader’s comprehension of the text is facilitated by the retrieval of information associated with a schema in long-term memory. When the reader
has inadequate prior knowledge to apply to a reading task, more cognitive effort will be required of the reader to assimilate the text to their prior knowledge and schemata (Richardson, Morgan, & Fleener, 2012; Woolley, 2011). Generally speaking, people read more literature, popular journalism, and simple expository texts than they do scientific or technical texts in their lives, thus non-specialist texts are usually easier to process for them than scientific texts (Alderson, 2000).

Usó-Juan (2006) tested 380 native Spanish-speaking undergraduates who exhibited a wide range of proficiency in English as a foreign language and knowledge of the academic topics being tested, and found using multiple regression analyses that English proficiency accounted for a range varying between 58% and 68% of EAP (i.e., English for academic purposes) reading, whereas discipline-related knowledge accounted for a range varying between 21% and 31%. This illustrates the importance of prior knowledge of a topic.

Text type

The first two categories and dimensions by RNAD mentioned above, i.e., discourse genre and discourse structure, can actually be discussed within the framework of text typology. The concepts “genre” “register” “style” “text type” “domain” are not easily distinguished. Different researchers from different backgrounds hold different opinions (for overviews, see Lee, 2001; Moessner, 2001). In this study, “text type” will be used as an overarching term.

In the field of translation studies, researchers have proposed different text classification methods (for an overview, see, e.g., Trosborg, 1997). For instance, Snell-
Hornby (1988) proposes a text categorization method based on the gestalt principle and the concept of prototypology. Her integrated approach diagram includes many levels. On the horizontal plane the diagram represents a spectrum, from literary translation (Bible, stage/film, lyric poetry, modern literature), to general language translation (newspaper/general information texts, advertising language), and to special language translation (legal language, economic language, medicine, science/technology). As the degree of variation in text difficulty within each text type is large, it is not possible for us to make such assertions as “modern literature is more difficult to translate than legal texts.” As a result, such subjective function-based text classifications are not particularly useful for discussions about text difficulty or translation difficulty.

By analyzing the co-occurrence distributions of 67 linguistic features in 481 spoken and written texts of contemporary British English using factor analysis, Biber (1988, 1989) develops a typology of texts in English with respect to a five-dimensional model of variation. The five dimensions are: 1) involved vs. informational production; 2) narrative vs. non-narrative concerns; 3) elaborated vs. situation-dependent reference; 4) overt expression of persuasion; 5) abstract vs. non-abstract style. Each dimension consists of a set of lexical and syntactic features that co-occur frequently in texts. In the dimension of “narrative vs. non-narrative concerns,” for instance, narrative concerns are marked by frequent instances of past-tense verbs, 3rd person pronouns, perfect-aspect verbs, public verbs, synthetic negation, and present-participial clauses; non-narrative concerns, whether expository, descriptive, or other, are marked by frequent instances of present-tense verbs and attributive adjectives. The relevance of Biber’s text typology to
translation difficulty research is that the linguistic features he used or dimensions (i.e., clusters of lexical and syntactic features) might cause difficulty in reading or translation. In the field of reading research, there have been some conclusions like “expository texts are harder to process than narrative texts,” and this might be “because of the greater variety of relationships among text units, possibly due to greater variety of content” (Alderson, 2000, p. 64). All these call for empirical studies.

2.2.1.2 Translation-specific difficulty

In the field of translation, Reiss (1982, from Hale & Campbell, 2002) proposes five textual aspects that give rise to text difficulty: the subject matter (semantic aspect), the register (material aspect), the type of language used (functional aspect), the pragmatics of the reader (pragmatic aspect) and the historical-cultural context (temporal, local or cultural aspect). Nord (1991) divides translation problems into four categories: 1) text-specific translation problems (e.g. a play on words), 2) pragmatic translation problems (e.g., the recipient orientation of a text), 3) cultural translation problems (e.g., text-type conventions), and 4) linguistic translation problems (e.g., the translation of the English gerund into German). Hill (1997, from Campbell & Hale, 1999) identifies five difficulty indicators, which she labeled “thematic, formal, stylistic, linguistic and syntactic” in her assessment of the reliability of French, German and Russian texts used for translation accreditation. Shreve, Danks, & Lacruz (2004) list the following factors: 1) textual or discourse variance 2) textual degradation including fragmentary texts and illegible texts 3) linguistic “distance” between the source text and the target text 4) cultural “distance” between the source culture and the target culture 5) lexical complexity
6) syntactic complexity 7) conceptual, topical or propositional complexity. The above delineations are all theoretical explorations.

Campbell and Hale are among the few researchers who have looked into translation difficulty empirically. They (1999) have identified several areas of difficulty in lexis and grammar, that is, words low in propositional content, complex noun phrases, abstractness, official terms, and passive verbs.

We can see that there is considerable overlap between these lists, and it seems that the three major aspects causing reading difficulty mentioned in previous section, that is, lexical and syntactic complexity, content and subject matter, and text type have also been covered.

What differentiates aspects of text difficulty from aspects of translation-specific difficulty can be narrowed down to one concept: equivalence. Equivalence is a key concept in translation theory (for an overview, see Kenny, 2009). It is “a relation of ‘equal value’ between a source-text segment and a target-text segment,” and “can be established on any linguistic level, from form to function” (Pym, 2010, p. 7). In her classic translation textbook In Other Words, Baker (1992, 2011) discusses equivalence at a series of levels: lexical (word level and above-word level), grammatical (e.g., number, gender, person, tense and aspect, and voice), textual (e.g., cohesion), and pragmatic (e.g., coherence and implicature). Of these levels, the lexical level probably creates the most noticeable difficulties for translators. For instance, if a translator mistranslates a term, it would be a glaring (and maybe unforgivable) mistake.

Kade (1968, from Pym, 2010, p. 29) proposes four types of lexical equivalence:
- One-to-one: One source-language item corresponds to one target-language item.

- One-to-several or several-to-one: An item in one language corresponds to several in the other language.

- One-to-part: Only partial equivalents are available, resulting in “approximate equivalence.”

- One-to-none: No equivalent is available in the target language.

One-to-one equivalence is common, especially for scientific and technical terms (e.g., airbag, motherboard). The major difficulty these specialized terms cause is that translators need to have access to such instruments as dictionaries if they are not familiar with the terms. One-to-several and one-to-part equivalences can create problems for translators. For instance, the word “cousin” has no exact equivalent in Chinese. Chinese distinguishes between older and younger cousins, between male and female cousins, and between cousins on the maternal side and on the paternal side. So, the word ‘cousin’ corresponds to 8 appellations in Chinese. If the word “cousin” is mentioned in the source text, the Chinese translator has to find the relevant details to ascertain which “cousin” it is referring to.

If a word has no equivalent in the target language, this can pose varying levels of difficulty. Baker (1992) lists 11 one-to-none equivalence situations:

1) Culture-specific concepts

2) The source-language concept is not lexicalized in the target language

3) The source-language word is semantically complex
4) The source and target languages make different distinctions in meaning
5) The target language lacks a superordinate
6) The target language lacks a specific term (hyponym)
7) Differences in physical or interpersonal perspective
8) Differences in expressive meaning
9) Differences in form
10) Differences in frequency and purpose of using specific forms
11) The use of loan words in the source text

For example, the word “grounded” as in “You are grounded.” is familiar to children in the USA. It means a child is “confined to his or her home outside school hours as a punishment” (Oxford English Dictionary), and not allowed to participate in social activities. As grounding children is not a discipline technique used by Chinese parents, it does not exist in Chinese culture. Therefore, it is not surprising that this word has no equivalent in Chinese language. To make things worse, this meaning of the word does not exist in most dictionaries (e.g., Collins English Dictionary, 2009), and it was added to the electronic version of the Oxford English Dictionary (second edition) as a draft partial entry only in September 2003. As a result, a Chinese translator would have considerable difficulty rendering this word.

Of course, there are strategies used by professional translators for dealing with various types of non-equivalence, such as translation by using a more general word, a more neutral/less expressive word, or cultural substitution (Baker, 1992). Nonetheless,
these non-equivalence, one-to-several equivalence, and one-to-part equivalence situations create difficulty for translators, especially for novice translators. The actual difficulty level of a translation task involves translator competence, which will be dealt with in the next section.

2.2.2 Translator factors

As mentioned in Section 2.1, the operator factors mainly include the individual’s cognitive capabilities, and past experience and training (i.e., prior knowledge). In the recent two decades, translation process researchers have been looking into these factors empirically (for an annotated bibliography, see Jääskeläinen, 2002), and these translator factors together are usually called “translation competence.” For instance, the EXPERTISE (Expert Probing through Empirical Research on Translation Processes) group, composed of leading European researchers in the field of translation processes, announced in 2002 that they tried to identify the translator’s knowledge base and cognitive underpinnings of expert behavior in translation. Their basic research objectives include translation aptitude, the development of translation competence, memory structures, monitoring operations in translation, creative mental processes in translation, etc.

Competence refers to personal qualities, skills and abilities, and exists in different degrees (Englund Dimitrova, 2005, p. 16). In order to investigate translation competence, researchers tend to break it down into a set of interrelated subcompetences, which can be studied in isolation or in combination with others (Schäffner & Adab, 2000, p. ix).
Many categories of translation subcompetences have been proposed (for an overview, see Kelly, 2005). For instance, Wilss (1976) suggests three subcompetences: 1) a receptive competence in the source language; 2) a productive competence in the target language; and 3) a supercompetence for transferring messages between linguistic and textual systems of the source culture and those of the target culture. Neubert (2000) proposes the following classification: 1) language competence, 2) textual competence, 3) subject competence, 4) cultural competence, and, 5) transfer competence.

The translation competence model proposed by the PACTE group (Process in the Acquisition of Translation Competence and Evaluation) from Universitat Autònoma de Barcelona in 2003 is perhaps the most influential. Based on empirical (mostly qualitative) studies, this model comprises five subcompetences: 1) bilingual subcompetence (i.e., knowledge in the two languages); 2) extra-linguistic subcompetence (i.e., knowledge about the source and target cultures, knowledge about the world in general, and subject knowledge); 3) knowledge about translation subcompetence (including knowledge about how translation functions and knowledge related to professional translation practice); 4) instrumental subcompetence (i.e., knowledge related to the use of documentation sources, and information and communication technologies applied to translation), and 5) strategic subcompetence (i.e., procedural knowledge for identifying translation problems and applying procedures to solve them). These subcompetences activate a series of psychophysiological mechanisms including cognitive components such as memory and attention, attitudinal aspects, and abilities such as creativity, logical reasoning, analysis and synthesis.
We can see that, except for transfer competence, the categories mentioned by Wilss and by Neubert are all included in the PACTE model. The PACTE group argues that the transfer subcompetence is not just one subcompetence because “[a]ll bilinguals possess a rudimentary transfer ability” (p. 57). They have replaced it with the strategic subcompetence, which occupies a dominant position among the subcompetences in the model. The strategic subcompetence is basically problem-solving ability, and it is widely accepted that problem solving (as well as decision making and reasoning) is an aspect of human intelligence (e.g., Gardner, 1983; Lohman & Lakin, 2011). In a word, these subcompetences are further categorizations of the individual’s cognitive capabilities and prior knowledge mentioned at the beginning of this section.

A basic approach in translation process research involves comparing translators of different competence levels, especially professional translators and translation students. Researchers tend to take professional translators as “experts” and translation students as “novices”, perhaps under the influence of Dreyfus & Dreyfus’s (1986) hierarchy of five learning stages (from novice stage to advanced beginner, competence, proficiency, and then to expertise stage). Experts are those who “are consistently able to exhibit superior performance for representative tasks in a domain” (Ericsson, 2006a, p. 3). Expertise in a domain does not necessarily transfer to another, although an individual with expertise in a specific domain may have useful knowledge of other domains (Shanteau & Pingenot, 2009). In translation studies, the concept of domain refers to task domain. It involves the translation of texts from a particular subject area and of a particular text type (Shreve, 2002), and relates to translator specialization.
Gouadec (2007) distinguishes two types of translation: general translation and specialized translation. Specialized translation refers to the translation of materials which belong to a highly specialized domain (e.g., law, medicine, literature), or the translation that involves the use of highly specific environments, tools and procedures (e.g., audio-visual translation, localization). General translation is the translation of materials that do not belong to any specific domain or particular type, such as letters, tourist brochures, press articles (in newspapers or magazines), user guides, etc. Based on these two types of translation, translators can be divided into generalist translators and specialist translators. Generalist translators translate materials which do not require a high degree of specialist or technical knowledge, while specialist translators usually focus on specific domains.

Setton & Guo (2009) conducted a survey among Chinese-English translators in Shanghai, Taipei and other Chinese cities. Forty-nine of the 62 respondents (that is, 79%) considered themselves generalist translators. A one-sentence poll\(^1\) held in 2010 on ProZ.com, which is probably the largest online portal for freelance translators and translation companies, showed that 86.2% of the 1,725 respondents specialized in certain field(s). Another poll\(^2\) held in 2011 on “In how many fields do you specialize (in your top language combination)?” indicated that 42.7% of the 1,731 respondents specialized in 2-3 fields, 33.3% of them in 4-5 fields. As people may have different ideas regarding what constitutes a specific field, these numbers should be taken with a grain of salt. However,

\(^1\) http://vls.proz.com/polls/8930
\(^2\) http://vls.proz.com/polls/6479
they did show that all translators have a few “comfort zones.” The real difference between generalist translators and specialist translators is that they draw on different translation subcompetences more often: generalist translators rely more on bilingual subcompetence than on extra-linguistic subcompetence (especially subject knowledge), while specialist translators draw on more subject knowledge than do generalist translators.

Naturally, different translation tasks require the use of certain subcompetences of translation more than others. For instance, if a source text contains a large number of terms unknown to the translator, then instrumental subcompetence (e.g., knowledge related to the use of the Internet and dictionaries) will be critical. On these subcompetences translators differ.

Since the mid-1980s, translation process researchers (e.g., Gerloff, 1986; Krings, 1986; Lörscher, 1986) have been comparing translators of different competence levels, and exploring aspects of translation competence. For lack of a theoretical framework in this field, these studies differ widely in their focus, scope and applicability to specific purposes (Fraser, 1996). Later, expertise theories were formally introduced into translation studies by Shreve (2002, 2006), who appeals for “leverage[ing] the expertise studies research to generate hypotheses or research questions for translation scholars to address” (2002, p.168). As a branch of research, expertise can be traced to de Groot (1965) and Newell & Simon (1972). During the past decades, expertise research has made rapid progress in many disciplines (see Ericsson, Charness, Feltovich, & Hoffman, 2006), especially cognitive psychology. The purpose of expertise research is to
understand the structure of expertise and the acquisition and maintenance of expertise (ibid.), and this corresponds to the purpose of translation competence research, which includes finding the components of translation competence and the process of translation competence acquisition (EXPERTISE, 2002; PACTE, 2003, 2005).

It is necessary to relate the findings in the expertise research field to the results of translation process studies. This may shed light on the characteristics of translation competence and its acquisition process. The following seven key characteristics enabling experts' successful performance are listed by Chi (2006).

1) Experts excel in generating the best solution and can do this faster and more accurately than non-experts.

In Tirkkonen-Condit's two-participant experiment (1987) involving one first-year student and one fifth-year student of translation, the mature student identified more problems but spent less time on the process, i.e. was more sensitized to potential problems but also more efficient in problem-solving. By contrast, Jääskeläinen (1999) found that the four non-professional translators spent less time on the task than did the four professionals. This contradiction may have originated from differences in choice of participants, translation materials (such as subject matter, text type), direction of translation, routine or non-routine work and other experimental design factors. Later in this study (see Section 4.3), the relationship between translation quality and time on task will be tested.

2) Experts can detect and see features that novices cannot, and can also perceive the “deep structure” of a problem or situation.
Tirkkonen-Condit (1992) found in her two-participant experiment that the professional's process was a top-down governed, holistic process which utilized textual knowledge, while the non-professional processed sentences in isolation and allowed her extra-textual knowledge to interfere with the contents of the source text. Jääskeläinen (1999) verified this point. In her experiment, the four professional translators relied more on textual and world knowledge in decision-making, while the four non-professionals remained at the linguistic surface level. It seems that this expertise rule applies to translation.

3) Experts spend a relatively great deal of time analyzing a problem qualitatively, developing a problem representation by adding many domain-specific and general constraints to the problems in their domains of expertise.

Lörscher (1996, p.31) found that professional translators mainly checked their utterances produced in target language with regard to their stylistic and text-type adequacy, while foreign language students only checked the solutions to their problems with respect to lexical equivalence and, to a lesser extent, to their syntactic correctness. Jääskeläinen (1999) discovered that non-professionals spent less time on the task than did professionals, and she attributed this result to novices' presupposed ignorance of potential translation problems. Therefore, this expertise rule also seems to be pertinent to translation.

4) Experts have more accurate self-monitoring skills in terms of their ability to detect errors and the status of their own comprehension.
Self-monitoring is closely related to revising in translation. Zhao’s (2004) study involving five expert translators and five translation students found that all the translators—both experts and students—went through all of the four progressive yet recursive stages (planning, comprehending, transferring, and self-monitoring), and the participants who did more high-level self-monitoring seemed to have produced better translations. This discovery agrees with the expertise rule.

5) Experts are more successful at choosing the appropriate strategies to use than novices.

Jääskeläinen (1993) distinguished translation strategies between global and local strategies, the former applying to the whole task, the latter to specific items (e.g. lexical searches), and concluded that global strategies were much more frequently used by professionals than by non-professionals. She also found that professionals appeared to follow their global strategies systematically through the task, whereas non-professionals seemed to proceed in a more haphazard way. Lörscher (1996) pointed out that professional translators and foreign language students were different in the distribution and frequency in the types of strategy.

6) Experts are more opportunistic than novices; they make use of whatever sources of information are available while solving problems and also exhibit more opportunism in using resources.

As EXPERTISE (2002) says, expert translators have intelligent use of information resources such as term banks, translation memory systems, or the Internet. Fraser (2000) found that in dealing with textual uncertainty, professionals and non-
professionals have some marked difference: professional translators show greater tolerance of ambiguity in the source text.

7) Experts can retrieve relevant domain knowledge and strategies with minimal cognitive effort, execute their skills with greater automaticity and exert greater cognitive control over those aspects of performance where control is desirable.

A number of studies (e.g., Jääskeläinen, 1999) reveal that routine translation conditions seem to result in higher levels of automatic processing (which is faster and more efficient than processing under conscious control) by professional translators, whereas non-routine conditions may prompt a less automatic behavior. Aside from the routine vs. non-routine condition differentiation, several empirical studies (e.g., Gerloff, 1988) show that “a higher degree of translational competence does not automatically correspond with a higher degree of translation process automatization” (Krings, 2001, p. 126). The reason may be that

although some aspects of the process do grow easier …, other aspects become concomitantly more complex. It is as if greater automaticity at one level (for example, at the level of basic linguistic decoding, e.g. identifying agreement between subject and verb or immediate comprehension of the usual meaning of a word) "frees up" processing capacity which may then be focused on other more complex levels of analysis. (Gerloff, 1988, p. 54)
Instead, some studies show that novice translators draw more intensively on automatic processes and make fewer conscious decisions (Alves & Gonçalves, 2007). Obviously, this question needs more research, and will be discussed in more detail in Chapter 5.

This section discussed components of translation competence and addressed differences between good translators and novice/poor translators by drawing on findings in translation process research and expertise studies. As translation competence develops, the difficulty level of a translation task changes for the translator.

2.3 Measuring translation difficulty

The preceding section dealt with potential sources of translation difficulty, which include translation factors (i.e., text difficulty and translation-specific difficulty) and translator factors. Accordingly, in order to measure translation difficulty, we need to measure text difficulty, recognize translation-specific difficulty (i.e., translation problems in a task), and assess translation difficulty (i.e., mental workload) for the translator.

2.3.1 Measuring text readability

To measure text difficulty, reading researchers have tended to focus on developing readability formulas since the early 1920s. A readability formula is an equation which combines the statistically measurable text features that best predict text difficulty, such as average sentence length in words or in syllables, average word length in characters, and percentage of difficult words (i.e., words with more than two syllables, or words not on a particular wordlist). Until the 1980s, more than 200 readability formulas had been published (Klare, 1984). Among them, the Flesch Reading Ease
formula (Flesch, 1948), the Dale-Chall formula (Dale & Chall, 1948), Gunning Fog Index (Gunning, 1952), the SMOG Formula (McLaughlin, 1969), the Flesch-Kincaid Readability test (Kincaid, Fishburne, Rogers, & Chissom, 1975), and the Fry Readability Formula (Fry, 1977) are the most popular and influential (Anagnostou & Weir, 2007; DuBay, 2004). These formulas use one to three factors with a view to easy manual application. Among these factors, vocabulary difficulty (or semantic factors) and sentence length (or syntactic factors) are the strongest indexes of readability (Chall & Dale, 1995; Kintsch & Miller, 1981). The following are the Flesch Reading Ease formula and the Flesch-Kincaid Readability formula.

Flesch Reading Ease formula:

\[
206.835 - 1.015 \left( \frac{\text{number of words}}{\text{number of sentences}} \right) - 84.6 \left( \frac{\text{number of syllables}}{\text{number of words}} \right)
\]

The resulting score ranges from 0 to 100; the lower the score, the more difficult to read the material. A score of 0 to 30 indicates the material is “very difficult” (graduate school level); 60 to 70, “standard”; 90 to 100, “very easy” (4th grade level) (Flesch, 1948, p. 230).

Flesch-Kincaid Readability formula:

\[
0.39 \left( \frac{\text{number of words}}{\text{number of sentences}} \right) + 11.8 \left( \frac{\text{number of syllables}}{\text{number of words}} \right) - 15.59
\]

The resulting number corresponds to a grade level in the USA. For instance, a score of 9.5 indicates that the material would be understandable by an average student in the 9th grade.
We can see that these two formulas adopt the same factors. The Flesch-Kincaid Readability formula simply translates the Flesch Reading Ease score into a U.S. grade level.

According to Klare (1984, from Anagnostou & Weir, 2007, p. 6), the steps to produce a readability formula are:

1) Choose (or develop) a large set of criterion-text passages of varied content and measured difficulty (typically, comprehension) values.
2) Identify a number of objectively stated language factors.
3) Count the occurrences of the language factors in the criterion passages and correlate them with the difficulty values of the criterion passages.
4) Select the factors with the highest correlations as potential indexes of readability.
5) Combine the indexes, using multiple regression techniques, into a formula.
6) [Optional] Cross-validate on a set of new criterion passages.

These steps will be followed in studying translation difficulty, and will be briefly explained below.

Obtaining the reading difficulty value of a passage

Readability formula developers’ most frequently used method for deriving the reading difficulty values of test passages is probably the cloze procedure developed by Taylor (1953). The cloze procedure works like this:

Test developers delete every fifth word from a passage and readers are asked to accurately guess the word that belongs in each blank. The readers’ rate of
accuracy in filling in the blanks estimates the text difficulty. Accuracy rate is based on readers identifying the exact word omitted from the passage. (Mesmer, 2008, p. 54)

The cloze score is the number of exact word replacements divided by the number of deleted words, expressed as a percentage. A passage that is assigned a difficulty value is called a criterion passage.

Subjective judgments are also used to measure reading difficulty of a passage. Carver (1975-1976) developed a new technique called the Rauding Scale that involves using the subjective ratings of qualified experts for estimating the difficulty of reading material. He claimed that this method was highly reliable and slightly more valid than some other readability formulas. Klare (1984) pointed out that reading difficulty assessment becomes more accurate as the number of qualified judges increases.

Chall, Bissex, Conard, & Harris-Sharples (1996) present a method that relies on a total impression rather than on an analysis of text features for the qualitative assessment of text difficulty. The method is based on matching a text to exemplars that have been scaled for comprehension difficulty in terms of vocabulary difficulty, sentence length and complexity, conceptual difficulty, idea density and difficulty. It is a kind of pairwise comparison. Their method for selecting passages as exemplars is as follows:

A large number of passages, more than were needed for the final scales, were selected. Their difficulty was assessed by a combination of several procedures: independent rankings of difficulty by the four researchers; independent rankings by several groups of teachers and reading specialists; comprehension difficulty as estimated by readability formulas; and also, for the expository social studies selections, cloze tests and students’ judgments of difficulty. The passages selected for the exemplars were those on which there was greatest agreement using the various difficulty measures. (p. 6)

**Identifying the language factors and counting their occurrences**
There are many language factors whose occurrences in a passage can be objectively counted. As mentioned earlier, Gray & Leary (1935) identified 289 factors that may be related to a book’s readability, and 64 of them are countable. Chall (1958, from Chall & Dale, 1995) noted that perhaps 100 or more text characteristics have been studied as possible indicators of reading difficulty in classic readability studies. Before the computer age, people identified the language features and manually counted their occurrences in a text.

Biber (1989, p. 7) analyzed the co-occurrence distributions of 67 linguistic features, which fall into 16 major grammatical categories: (A) tense and aspect markers, (B) place and time adverbials, (C) pronouns and pro-verbs, (D) questions, (E) nominal forms, (F) passives, (G) Stative forms, (H) subordination features, (I) prepositional phrases, adjectives, and adverbs, (J) lexical specificity, (K) lexical classes, (L) modals, (M) specialized verb classes, (N) reduced forms and dispreferred structures, (O) coordination, and (P) negation. These features can be identified automatically by his computer programs.

There are many text analysis tools currently available for analyzing various aspects of a text, e.g., syntax, semantics, propositions, text cohesion and coherence, and text type (see Castello, 2008; Graesser, McNamara, & Louwse, 2011). If some language features cannot be identified by computers, manual annotation of these features can facilitate the automatic analysis.

Combining the indexes into a formula
This involves a statistical technique called multiple regression analysis, which is one of the most widely used statistical techniques in the social sciences. Multiple regression is for investigating the joint effect of one or more predictor variables (e.g., average sentence length, average word length) on a response variable (e.g., readability). It can determine which predictor variables are important predictors of the response variable, and predict the value of the response variable from the predictor variables (see e.g., Ott & Longnecker, 2010).

Take Bonnuth’s study in 1964 as an example (from Chall & Dale, 1995). He tested and correlated 47 readability variables with his criterion passages, and found that the linguistic factors other than traditional measures of readability (e.g., sentence length, word length) added little to the overall predictions of reading difficulty. As a result, only three traditional measures of readability (i.e., average word length in letters, words not on the Dale-Chall 3,000 Simple Word List, and average sentence length) were retained in his readability formula. Compared with the sentence factors, the word factors are more predictive (Chall & Dale, 1995, p. 63).

It should be noted that multiple regression analysis can reveal how variables are related to each other, and whether some variables can predict a response variable, but cannot prove causal relations among variables (Urdan, 2010, p. 156). That is, the readability formula does not necessarily imply causal connections between the chosen predictor variables (e.g., vocabulary difficulty, sentence length) and the response variable (i.e., readability). Reading difficulty often comes from the ideas rather than the words or sentences. However, difficult passages that express difficult, abstract ideas tend to
contain hard words, while easy passages dealing with familiar, concrete ideas tend to use familiar words (Chall & Dale, 1995; Rayner & Pollatsek, 1989, p. 319). In addition, according to Zipf’s law (Zipf, 1935), word frequency and word length are inversely related, i.e., short words occur with high frequency while longer words occur with lower frequency as a result of a biological principle of least effort. These might explain why readability formulas work.

Over the decades readability formulas have been proved to be valid and fairly reliable (e.g., Fry, 1989; Klare, 1974-1975), and have been adopted by publishers, schools, and software producers (e.g., the Flesch Reading Ease formula and Flesch-Kincaid Readability formula are used by Microsoft Word).

Yet, the readability formulas mentioned so far are all for English native speakers (especially people in the USA). In the professional world, translators are usually expected to work from a foreign language into the mother tongue (Lonsdale, 2009), and English has been the most translated language (Sin, 2011). That is, most translators translate from English into their mother tongue. The pertinent question is whether the mainstream readability formulas work for EFL (English as a Foreign Language) or ESL (English as a Second Language) students.

Carrell (1991) investigated the first and second language reading comprehension of adult native speakers of Spanish and English who were foreign or second language learners of the other language, and found that reading in a second language is a function of both first language reading ability and second language proficiency. Nonetheless,
readability formulas predict difficulty rather than explain its causes, and their suitability for EFL/ESL readers can only be determined by empirical studies (Greenfield, 2004).

Brown (1998) administered cloze tests to 2,298 Japanese EFL university students, compared the observed average cloze scores on the passages with scores predicted by six readability formulas (including Flesch Reading Ease, Flesch-Kincaid Readability, and others), and concluded that first language readability indices were only weakly related to EFL Difficulty. However, he noted that a readability index that is more highly related to the performance of second language learners can be created, and EFL/ESL readability might best be estimated separately for students from different language backgrounds.

This section briefly discussed readability formulas and their development process, which has significant implications for developing a translation difficulty formula. In addition, as mentioned earlier, there is a possibility that readability formulas can be used to measure translation difficulty, and this will be tested in this study.

2.3.2 Recognizing translation problems and difficulty

There are many ways to locate translation difficulty “points” in a source text for a translator. The three methods common in translation studies are translation quality assessment (or more specifically, grading translations), analyzing verbal protocols, and recording and analyzing translation behavior, which will be detailed in the following sections. But first, it is necessary to review the method used by Campbell and Hale and by Jensen (2009), who are among the few researchers who have investigated translation difficulty empirically.
The way Hale & Campbell (2002) assess the difficulty of a source text is to count the number of different renditions in a group of translators translating that text. Their rationale is “the different renditions represent the options available to the group of subjects, and that each subject is faced with making a selection from those options;” “where there are numerous options, each subject exerts relatively large cognitive effort in making a selection; where there are few options, each subject exerts relatively small cognitive effort” (p. 15). In a way, this contradicts Pym’s (2003) opinion on translation competence. He identifies two components of translation competence, namely, 1) the ability to generate a series of more than one viable target text for a pertinent source text; 2) the ability to select only one viable target text from this series, quickly and with justified confidence (p. 489). In other words, poor translators generate fewer translation equivalents than do good translators. In line with Hale & Campbell’s logic, this means poor translators exert relatively small cognitive effort than do good translators. Obviously, this does not make much sense.

In fact, both of the two theories are problematic. For instance, if a proper name has only one equivalent in the target language, and it takes the translators a lot of time to find that equivalent, is this easy or difficult? According to Hale & Campbell’s standard, finding that equivalent should be classified as easy, as there is only one rendition. However, it takes a lot of cognitive effort and should be regarded as difficult.

Jensen (2009) employed three indicators of translation difficulty in his study: readability indices, word frequency, and non-literalness (that is, the number of occurrences of non-literal expressions, i.e. idioms, metaphors, and metonyms). His
reasoning was that “If one employs several indicators that all point in the same direction, it is reasonable to assume that they can be used as a reliable measure” (p. 76). But, what if three indicators do not point in the same direction? For instance, text A is more complex than text B in terms of readability indices and word frequency, but has fewer non-literal expressions than text B. Can we conclude that text A is more difficult to comprehend or translate than text B? No. As mentioned in Section 3, there are actually more variables at play (such as domain or text type). Another problem with his approach is that word frequency is a factor that most readability formulas take into account.

2.3.2.1 Grading translations

There are two basic approaches to scoring in language testing: holistic and analytic (e.g., Hughes, 2003; McNamara, 2000). According to a survey conducted by Waddington (2001) among 52 translation teachers from 20 European and Canadian universities, 38.5% used a holistic method, 36.5% used an analytic method (based on error analysis), and 23% combined the two.

The holistic method is based on the idea that the whole is greater than the sum of its parts (Johnson, Penny, & Gordon, 2009). Graders are usually required to provide a single grade for a translation, and they have to combine all the prominent features of a translation to arrive at an overall judgment of its quality. They normally need to refer to a holistic rubric so that the grading can be more systematic and objective. A holistic rubric defines performance criteria and levels but does not indicate specific components of the performance (Gareis & Grant, 2008). For instance, the grading rubric developed by American Translators Association (ATA) (version 2009) contains four levels, as follows:
- **Strong**: Translated text conveys meaning fully and accurately as specified by Translation Instructions.

- **Acceptable**: Translated text conveys meaning well enough to be useful to intended reader; occasional mistranslations, omissions or additions may slightly obscure meaning.

- **Deficient**: Translated text does not convey meaning well enough to be useful to the intended reader; mistranslations, omissions or additions may obscure meaning.

- **Minimal**: Translated text would be nearly useless to intended reader; frequent and / or serious mistranslations, omissions or additions obscure or change meaning.

Of course, a rubric can have three levels (unacceptable, barely acceptable, clearly acceptable), or six levels (unacceptable translation, inadequate translation, barely adequate translation, competent translation, very competent translation, and outstanding translation), or other number of levels suitable for the purpose it is intended to be used (Angelelli, 2009, p. 39).

The analytic method entails that graders assign scores for different components or characteristics of the task, and then add up these scores to obtain an overall score (Sullivan, 2009). In the field of translation, the analytic rubric is usually an error classification scheme. An error, by severity, can be a major one or a minor one, so a weight in the form of a numerical value can be assigned to each error. Currently there are
quite a few translation error classification schemes, e.g., SAE J2450, BlackJack, Canadian Language Quality Measurement System (Sical), ATA framework for standardized error marking, LISA (Localization Industry Standards Association) QA Model, MeLLANGE, and others (see, e.g., Dunne, 2009; Secară, 2005; Williams, 2004). Take SAE J2450 (SAE, 2001), originally designed for the automotive industry, as an example. The following are its error categories and points to be deducted.

### Table 2.1 SAE J2450 Translation Quality Metric

<table>
<thead>
<tr>
<th>Error category</th>
<th>Weight: Serious</th>
<th>Weight: Minor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wrong Term</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Syntactic Error</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Omission</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Word structure or agreement error</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Misspelling</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Punctuation Error</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Miscellaneous Error</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

The primary advantages of holistic scoring are the simultaneous consideration of all components of the response and time efficiency, while the strength of analytic scoring lies in assessing the examinee’s specific strengths and weaknesses and identifying the particular components of competence in a field (Welch, 2006). A commonly cited disadvantage of analytic scoring is that it tends to reduce and oversimplify the components of competence in a field (ibid.). For instance, SAE J2450 does not have a “style” category.
When comparing the two scoring methods, some researchers (e.g., Knoch, 2009) think that the analytic method is more reliable and less subjective than the holistic method. Other researchers, on the contrary, believe that “[t]he more holistic the rubric, the fewer the gradations and shades of gray and thereby, the more objective and reliable the scores can be” (Wormeli, 2006, p. 46). In a translation study, Waddington (2001) hypothesized that methods of assessment based on error analysis were more reliable than holistic methods, but this hypothesis was not verified.

During the last few decades, translation researchers have become increasingly impatient with the holistic method (also called “subjective impressionism”), and have been trying to introduce objectivity in judging translation quality (Al-Qinai, 2000). However, measuring translation quality is like judging a beauty contest; it is an inherently subjective process and relies on human judgment. Even if it were an objective process and all translation errors were mechanical (e.g., capitalization, punctuation, single word errors), graders still would not assign the same score to the same translation. Fearn (1982, from Speck, 2009, p. 147) is a pertinent example. He asked 33 teachers to evaluate the mechanical accuracy of one elementary school writing, and the teachers found “anywhere from no punctuation errors to nine punctuation errors.” In this light, the subjectivity versus objectivity debate would not be productive. The real issue is how to minimize bias and increase the reliability and validity of the grading methods.

There are many types of grader (or rater) bias, such as central tendency (i.e., assigning scores primarily around the scale midpoint), fatigue, handwriting, leniency/severity, skimming (i.e., not read the whole response), trait (i.e., focusing on one
aspect), and others (Johnson, Penny, & Gordon, 2009, p. 211). For instance, Sweedler-Brown (1992, from Speck, 2009) compared the holistic scores for 27 essays represented in three graphic modes (i.e., typed, nicely handwritten, and poorly handwritten) and evaluated among about 700 other essays, and found that both typed and poorly handwritten essays received lower scores than nicely handwritten essays even when graders were explicitly trained to be aware of the handwriting bias. This means that the graphic appearance of a translation may have an effect on holistic scoring of that translation. Schaefer (2008) explored the rater bias patterns of native English-speaker raters when they rate EFL essays, and found that “there tended to be more severe or lenient bias towards higher ability writers than lower ability writers,” and some raters “rated higher ability writers more severely and lower ability writers more leniently than expected” (p. 465). We can see that this leniency/severity bias is related to the central tendency bias.

These biases need to be minimized via developing scoring rubrics, grader training, identifying benchmark responses that provide concrete examples of each proficiency level or exemplify each point of the scoring rubric, and other measures (Johnson, Penny, & Gordon, 2009; Linn, 2006).

Increasing the number of graders is a useful way for improving grading reliability. In a study evaluating second language learners' writing skills, Brown & Bailey (1994) found that three independent raters of each student's paper increased interrater reliability compared to two independent raters of each student's paper.
When evaluating translations, the grader first establishes an ideal translation and then compares it with the examinees' translations (Lauscher, 2000). As different graders have different ideas about what an ideal translation would be like, it is necessary for them to produce a model translation together. As Lauscher (ibid.) says, “translation quality is ultimately a matter of agreement and consensus” (p. 149).

Each of these ideas mentioned in this section has informed this study in grading translations. For instance, both holistic and analytic methods would be used; three graders rather than only two would be employed; a model translation would be produced for each source text by the graders.

2.3.2.2 Analyzing verbal protocols

The theory that verbal protocols can be used to elicit data on cognitive processes was proposed by Ericsson and Simon (1980, 1993), and they have provided substantial empirical support for it. Ericsson and Simon hold that participants can generate verbalizations, subordinated to task-driven cognitive processes, without changing the sequence of their thoughts (1993, p. xxxii). Ever since, analyzing think-aloud protocols (TAP) has become one of the core research methods in expertise studies (e.g., Ericsson, 2006b) and problem solving (e.g., Robertson, 2001). In this kind of research, participants are requested to speak out their thoughts while performing a task (e.g., translating a text). This way, verbal reports provide explicit information about the knowledge and information heeded in solving a problem, and from the verbal reports, researchers can infer the processes (Simon & Kaplan, 1989, p. 23). For instance, if a participant thinks
aloud that “The word puzzles me.” or “What does this mean?”, the researcher would know that this translator encounters a problem.

Analyzing think-aloud protocols has been a major research method in translation process research since the 1980s (for an annotated bibliography, see Jääskeläinen, 2002). Almost all of the research findings regarding translation process mentioned in Section 2.2.2 were made in studies employing this method. In recent years, there have been some controversies as to the suitability of verbal protocol analysis for translation process research in terms of its validity and reliability. Although empirical studies are needed to solve some of the doubts and questions, to date, no study has falsified this method for translation process research (see Sun, forthcoming).

2.3.2.3 Recording and analyzing translation behavior

Verbal protocols are used to look into thoughts and their sequences, and they are verbal data. By comparison, recording methods are used to record overt behavior precisely. Their data often can be transformed into numbers and used in correlational analysis. The oft-used recording methods in translation research include keystroke logging, screen recording, and eye tracking.

A keystroke logging software tool (e.g., Translog, ScriptLog, Inputlog) records all the keyboard activity, and generates a lot of recorded data consisting of information concerning pausing (where and when pauses occurred, and for how long) and the history of all keyboard actions and cursor movements; it is used to study pause location, pausing in relation to planning and discourse production, and revision behavior (see Sullivan & Lindgren, 2006). If a participant pauses for a relatively long time during translation, the
researcher would know where this translator has a problem (e.g., Jensen, 2009). One characteristic of keystroke logging is that only the writing process involved in translating is recorded.

A keystroke logger can only record keyboard actions and cursor movements within its interface. To track the translator’s behavior outside the keystroke logger (e.g., consulting an electronic or online dictionary), screen recording is often used. Such a tool (e.g., CamStudio, Camtasia) can record all screen and audio activity on a computer and create AVI video files. Eye tracking can measure eye movements including the number of fixations, fixation durations, attentional switching, and scanpath similarity (see Duchowski, 2007).

The fundamental assumption in using translation behavior data is that there is a correlation between outside behavior and cognitive processing, and in most cases, it is only possible to speculate on what cognitive processes a translator is engaged in if only one of the methods is used (Jakobsen, 2011, p. 37). For instance, if a participant’s mind wanders during the translation process, pause data produced by keystroke logging will be misleading. In order to maximize the chance of reconstructing the translation cognitive processes, researchers tend to use multiple methods. For instance, Shreve, Lacruz, & Angelone (2010) combine keystroke logging and eye tracking in a study designed to investigate the effects of syntactic difficulty on cognitive effort in sight translation and cognitive differences between written and sight translation.

As verbal protocol analysis and recording and analyzing translation behavior are usually time-consuming and labor-intensive, they are usually suitable for studies
involving a few participants and/or relatively short texts. For instance, transcribing an hour of speech takes over ten hours. In Krings’ study (2001, p. 213), transcribing about 100 hours' VCR recordings took about 1600 working hours. A study for developing a readability (or translation) formula, however, needs a large number of participants. For this reason, of the three methods, only translation quality assessment (i.e., grading translations) was used in this study. The translation grade can be taken as an indicator of translation performance, and translation errors can show where the translation difficulty “points” lie.

2.3.3 Measuring mental workload

The previous section discussed about how to locate and diagnose difficulty “points” in a translation. This section addresses how to measure the global difficulty level of a translation for a translator.

There are a large number of techniques for measuring mental workload, and they can be classified into three major categories: 1) subjective measures, 2) performance measures, and 3) physiological measures (O'Donnell & Eggemeier, 1986).

2.3.3.1 Subjective measures

Subjective measures typically involve having participants judge and report their own experience of the workload imposed by performing a specific task. The rationale for using subjective measures is that increased capacity expenditure will be associated with subjective feelings of effort or exertion that can be reported accurately by the participant
unless the workload is beyond what the participant can handle (O’Donnell & Eggemeier, 1986, p. 7).

Subjective measures have been very frequently used to evaluate mental workload mainly because of their high face validity and easy implementation (Vidulich, 1988), and the most commonly used subjective measure is the rating scale. Over the decades, rating scales have proven to be valid, sensitive, non-intrusive measures of workload that enjoy a high degree of acceptance among the participants and require minimal instrumentation (Wilson & Eggemeier, 2006).

Among the subjective rating scales designated specifically for assessing workload, the most frequently employed and evaluated with respect to their reliability, validity and sensitivity to change in workload are the NASA task load index (TLX), the subjective workload assessment technique (SWAT), and the Cooper-Harper scale (MCH) (Wilson & Eggemeier, 2006).

Based on 16 different experiments with different kinds of tasks, Hart & Staveland (1988) developed NASA-TLX (Task Load Index), which includes six workload-related subscales, as follows:

1) *Mental Demand:* How much mental and perceptual activity was required (e.g., thinking, deciding, remembering, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?

2) *Physical Demand:* How much physical activity was required (e.g., pushing, pulling, etc.)?
3) **Temporal Demand**: How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred?

4) **Effort**: How hard did you have to work (mentally and physically) to accomplish your level of performance?

5) **Performance**: How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

6) **Frustration Level**: How insecure, discouraged, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Each subscale is presented as a line divided into 20 equal intervals anchored by bipolar descriptors (e.g., low/high, good/poor), as shown in Figure 2.1:

![Fig. 2.1 A NASA-TLX subscale](image)

As the six factors may contribute different loads to the workload of the specific task, participants are required to give a weight to each factor through pairwise comparison of the relative importance of each factor. More weight is given to ratings of factors deemed more important in creating the workload of a task in order to enhance the
scale’s sensitivity. If a factor is considered not contributing to the workload, it can receive a weight of 0.

Ever since its introduction, NASA-TLX has been subjected to a number of independent evaluations with respect to its reliability, validity, sensitivity, and utility (Hart, 2006). It has been adopted to evaluate the workload of tasks with focus on manual control, perception, short-term memory, cognitive processing, or parallel and serial dual-tasks; it is now “being used as a benchmark against which the efficacy of other measures, theories, or models [of workload measurement] are judged” (ibid, p. 907).

However, it should be noted that there are several studies (e.g., Nygren, 1991; Pfendler, 1991, from Embrey, Blackett, Marsden, & Peachey, 2006) showing that the weighting scheme is unnecessary. For instance, Moroney, Biers, Eggemeier, & Mitchell (1992) found that the weighted and unweighted scores were highly correlated ($r=0.94$) and analyses revealed no differences between the weighted procedure and unweighted procedure.

In the field of cognitive load theory, a well-known and extensively used subjective rating scale is a 9-point Likert scale first used by Paas (1992). This scale ranges from “very, very low mental effort” (1) to “very, very high mental effort” (9). Later studies show that this simple subjective rating scale has been shown to be a very “sensitive measure available to differentiate the cognitive load imposed by different instructional procedures” (Sweller, Ayres, & Kalyuga, 2011, p. 74).

For measuring the difficulty of tasks in second language instruction, Robinson (2001) designed a Likert rating scale questionnaire, which includes five items:
perceptions of task difficulty, anxiety, interest level, motivation, and perceptions of ability. Ishikawa (2011) added three more items to this task-difficulty questionnaire: concentration, response to time pressure, and anxiety.

As noted earlier, mental workload is generally believed to be a multidimensional construct. Therefore, NASA-TLX was used in this study to evaluate the participants’ subject workload during translation. In addition, a likert scale like the one used by Paas was employed for the participants to predict the translation difficulty level of a text. Obviously, NASA-TLX is not appropriate for predicting mental workload. For instance, asking the participants to predict their Effort before they actually do the translation would not be reliable.

2.3.3.2 Performance measures

Performance measures derive an index of workload from some aspect of the participant’s behavior or activity (O'Donnell & Eggemeier, 1986). It is generally assumed that low to moderate levels of workload or information processing demand are associated with acceptable levels of operator performance, while high levels of workload are associated with degraded levels of operator performance (Wilson & Eggemeier, 2006, p. 814). In other words, the relationship between workload and performance is not linear. This has been realized by researchers in the field of human factors and ergonomics since the 1970s (e.g., Meister, 1976).

Figure 2.2 depicts a hypothetical relationship between workload and performance (from O'Donnell & Eggemeier, 1986, p. 3). For easy tasks, increased workload is not accompanied by variations in performance, or may lead to improved performance; for
moderately difficult tasks, participants may not be able to increase their effort enough to meet the task demands, and thus increases in workload lead to gradually declined performance; for very difficult tasks that participants perceive as unreasonable, they reduce their effort, bring the workload to normal levels, and their performance deteriorates (Charlton, 2002; O'Donnell & Eggemeier, 1986). That is, task underload may lead to artificially enhanced performance, whereas task overload may result in a “floor effect” (O'Donnell & Eggemeier, 1986). In an experiment utilizing functional near-infrared spectroscopy (fNIRs) to examine the relationship between mental workload, level of expertise, and task performance, Bunce et al. (2011) found that as the level of taskload in a video game moved from moderate to high, the four participants with high practice performed better, while the four novices’ performance dropped precipitously.

Fig. 2.2 Hypothetical relationship between workload and performance

There are two major categories of performance measures: primary task measures and secondary task methodology. The primary task method assesses the capability of the
participant to perform the task of primary interest (e.g., translating a text) (Wilson & Eggemeier, 2006). For tasks that present no exceedingly low or high workload for the participants, primary task measures will differentiate the levels of workload. Because performance measures are insensitive to very low workload levels (i.e., very easy tasks), adding a secondary task (e.g., memory or mental arithmetic tasks) that is performed concurrently with the primary task would increase the workload to a moderate level and make performance measures sensitive (Eggemeier & Wilson, 1991). The secondary task should compete for the same processing resources as does the primary task, and participants are usually instructed to maintain error-free performance on the primary task at the expense of the secondary task, so that performance of the secondary task can be used to derive an estimate of primary task workload (O'Donnell & Eggemeier, 1986; Tsang, 2006).

Two commonly used workload indicators are speed (i.e., time-on-task) and accuracy (i.e., number of errors). They have been proved to show sensitivity to workload manipulations (see O'Donnell & Eggemeier, 1986, p. 21).

In this study, both of the two workload indicators were measured, and the relationship between workload and performance were tested.

2.3.3.3 Physiological measures

Physiological measures detect changes in the participant’s body for inferring the level of workload related to a task. There are various classification methods for them (e.g., Tsang, 2006; Vice, Lathan, Lockerd, & Hitt, 2007). For instance, O'Donnell & Eggemeier (1986) divide physiological measures into four categories, as follows:
- Measures of Brain Function: e.g., electroencephalogram (EEG)
- Measure of Eye Function: e.g., corneal reflex, electrooculogram (EOG), and pupil response
- Measures of Cardiac Function: e.g., electrocardiogram (EKG), blood pressure, heart rate, and oxygen concentration
- Measures of Muscle Function: e.g., electromyogram (EMG)

Compared with subjective and performance measures, physiological measures are typically “online, covert, and continuous,” and “can be linked in time to the expected occurrence of the psychological state or process being indexed, thereby providing simultaneous evidence of the strength or operation of the state or process” (Blascovich, 2004, p. 881). However, they are affected by environmental influences and the participant’s physical state, and compared with those reflective of bodily functions, the physiological signals that are indicative of mental demands could be small (Tsang, 2006). When collecting data using physiological measures, researchers need to consider such factors as invasiveness, cost, sensitivity, validity, practicality, and reliability (Vice, Lathan, Lockerd, & Hitt, 2007).

Each of the three kinds of techniques for measuring mental workload discussed in this section can provide valuable information, and they have different strengths and weaknesses. Because mental workload is multidimensional and real-world tasks may involve more than one component, researchers usually use multiple workload measures in
order to obtain a more complete picture of the mental workload (Tsang, 2006). In this study, subjective measures and performance measures were used.

Sometimes, conflicting results from multiple measures arise. For instance, several studies in the field of cognitive load theory found a cognitive load difference based on subjective measures but no group treatment effect on performance measures (see Sweller, Ayres, & Kalyuga, 2011, p. 75). In this case, we can follow Jex’s (1988, p. 14) suggestion that “the fundamental measure, against which all objective measures must be calibrated, is the individual's subjective workload evaluation in each task.”

This chapter explores sources of translation difficulty and ways for measuring them. It draws on many research fields besides translation studies, and helps form a foundation for this line of inquiry. The steps for developing readability formulas will be followed to develop a translation difficulty formula. Mental workload theories from the field of human factors and ergonomics help locate sources of translation difficulty and suggest way for measuring them. The methods adopted in this study, which will be detailed in next chapter, were based on this chapter’s explorations.
CHAPTER 3

Methods

The purpose of this study was to find a method to measure difficulty in English-Chinese translation. In order to accomplish this purpose, it was necessary to measure text difficulty, recognize translation-specific difficulty (i.e., translation problems in a task), and assess translation difficulty (i.e., mental workload) for the translator. Readability formulas are based on statistically measurable factors such as average sentence length and percentage of difficult words. Long sentences and difficult words, as translators would agree, increase the level of translation difficulty. Therefore, readability formulas might make it possible to measure translation difficulty. If they were proved to be ineffective, a model would be developed to predict the translation difficulty level of a text.

This chapter discusses the methodology that was used to answer the research questions. It consists of a description of participants, materials, test procedure, pilot study, translation grading, and normality checking.

3.1 Participants

Forty-nine third-year undergraduate students in translation/English from a university in Shandong Province, China, and 53 first-year graduate students in translation from a university in Beijing and that university in Shandong Province participated in this study during May and September 2011. One undergraduate student participated in only
one of the two sessions. Each group was homogeneous in terms of age and years of English learning and translation training experience. None of them worked as professional translators. They all spoke Mandarin Chinese as their first language, started learning English as a foreign language from the 6th grade, and had had classes in translation theory and practice. Table 3.1 summarizes the participants’ information.

Table 3.1 Participants in this study

<table>
<thead>
<tr>
<th>Participants</th>
<th>University 1</th>
<th>University 2</th>
<th>Total</th>
<th>Age³</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd-year Undergrad Students</td>
<td>49</td>
<td>0</td>
<td>49</td>
<td>Around 20</td>
</tr>
<tr>
<td>1st-year Graduate Students</td>
<td>12</td>
<td>41</td>
<td>53</td>
<td>Around 22</td>
</tr>
</tbody>
</table>

All participants took part in this study on a voluntary basis. With their permission, identifiable information was collected in order to provide feedback on their translations to them. Their identifiable information would never be used for other purposes. Each participant received a pen, correction tape, and a notebook or stopwatch as remuneration for their time.

3.2 Materials

The test materials included short passages to be translated from English into Chinese, and pre-translation and post-translation ratings.

³ The two ages are approximations; no data were collected in this regard.
3.2.1 Test passages

Fifteen passages were used in this study. They were all informational and expository texts. Most of them were from eHow.com and About.com. Their lengths ranged from 121 to 134 words. See Appendix A for the passages.

The reason for using informational texts was threefold. First, informational texts “comprise the majority of reading and writing done by adults” (Hoyt, Mooney, & Parkes, 2003, p. 1). Second, it is difficult to find self-contained short narrative or literary passages. Third, taking into account the text type variable would require a larger sample of text passages in this study, and thus all texts were of the same text type.

An important consideration when selecting passages for a pool of prospects was that no substantial knowledge in a specialized field would be required from the participants. The texts should be self-contained and required no additional context. They should not be very long so that participants could translate six of the passages in about 2.5 hours. In addition, no Chinese translations of the English texts could be found by the participants on the Internet (although they were not allowed to access the Internet during the test).

Fifty passages in the pool were measured using the Flesch Reading Ease formula. Depending on their Flesch Reading Ease scores, passages in the scope of 70-80 belonged to the Easy group; those in the scope of 46-55, Medium group; those in the scope of 20-30, Difficult group. The three groups would be compared in later analysis. As measured by the Flesch–Kincaid Grade Level, a score of 85 corresponds to a 5th grade level, while below 30 is graduate school level.
In order to locate the subject matter and terminology challenges that participants might not be able to handle even with a general English-Chinese dictionary, this author translated the prospective texts in the three reading difficulty groups, and narrowed down the number of eligible passages to 15.

The following is the list of test passages and their Flesch Reading Ease scores.

Table 3.2 Test passages and their Flesch Reading Ease scores

<table>
<thead>
<tr>
<th>Number</th>
<th>Category</th>
<th>Passage</th>
<th>Flesch Reading Ease Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Difficult</td>
<td>Impacts of Ecosystems</td>
<td>21</td>
</tr>
<tr>
<td>D2</td>
<td>Difficult</td>
<td>Natural Cures for Tiredness</td>
<td>28</td>
</tr>
<tr>
<td>D3</td>
<td>Difficult</td>
<td>What Determines Health</td>
<td>29</td>
</tr>
<tr>
<td>D4</td>
<td>Difficult</td>
<td>Superfoods</td>
<td>29</td>
</tr>
<tr>
<td>D5</td>
<td>Difficult</td>
<td>Disruptions to Ecosystems</td>
<td>30</td>
</tr>
<tr>
<td>M1</td>
<td>Medium</td>
<td>Why Men Snore</td>
<td>46</td>
</tr>
<tr>
<td>M2</td>
<td>Medium</td>
<td>Weight Control</td>
<td>48</td>
</tr>
<tr>
<td>M3</td>
<td>Medium</td>
<td>Friendships</td>
<td>49</td>
</tr>
<tr>
<td>M4</td>
<td>Medium</td>
<td>How to Get a Good Memory</td>
<td>51</td>
</tr>
<tr>
<td>M5</td>
<td>Medium</td>
<td>Sunburn</td>
<td>53</td>
</tr>
<tr>
<td>E1</td>
<td>Easy</td>
<td>How to Avoid Germs in Public</td>
<td>72</td>
</tr>
<tr>
<td>E2</td>
<td>Easy</td>
<td>Breathing Exercises</td>
<td>74</td>
</tr>
<tr>
<td>E3</td>
<td>Easy</td>
<td>How to Teach a Child Table Manners</td>
<td>76</td>
</tr>
<tr>
<td>E4</td>
<td>Easy</td>
<td>How to Sneeze Without Spreading Germs</td>
<td>79</td>
</tr>
<tr>
<td>E5</td>
<td>Easy</td>
<td>Running</td>
<td>80</td>
</tr>
</tbody>
</table>

A software program called Readability Studio (version 3.2.7) was used to analyze the readability of the 15 English source texts. This program contains many popular
readability formulas, such as Flesch Reading Ease, Flesch-Kincaid, New Dale-Chall, and SMOG. It can calculate the number of sentences, average sentence length, number of hard words, etc. Figure 3.1 presents a more complete list of the categories.

![Fig. 3.1 Readability Studio](image)

### 3.2.2 Translation ratings

The pre-translation rating asked participants to predict the translation difficulty level of the passage they just read but had not started to translate on a 0-10 likert scale with 0 being extremely easy and 10 being extremely difficult (as shown in Figure 3.2).
The post-translation ratings had two parts. The first part was a short survey, addressing how much time they spent on the translation, which of the two (i.e., source text comprehension and reverbalization in the target language) was more difficult for them, which aspects in source text comprehension (i.e., lack of background knowledge, too many difficult words, sentences too long, and others) created more problems for them, and which aspects in reverbalization (i.e., words, sentences, style, and others) created more problems for them. For the details see Appendix B.

The second part was adapted from NASA-TLX to evaluate subjective workload. Four categories were used, including Mental Demand (i.e., was the task easy or demanding, simple or complex), Effort (i.e., how hard did you have to work to accomplish your level of performance), Frustration level (i.e., how insecure, discouraged, irritated, and stressed did you feel during the task), and Performance (i.e., how satisfied were you with your performance). The categories “physical demand” and “temporal demand” were not used as they were not applicable, or they were just given a weight of 0. See Appendix C for the details.
3.3  Design and procedure

3.3.1  Experimental design

Each participant translated 6 passages (2 Easy, 2 Medium, and 2 Difficult) on paper in two sessions with a week’s interval between them. In each session they translated 3 passages (one Easy, one Medium, and one Difficult). The order in which they translated the passages was counterbalanced using a Latin square design. Tables 3.3 and 3.4 present part of the test passage order arrangement:

Table 3.3 Part of the test passage order arrangement

<table>
<thead>
<tr>
<th>Undergraduate Students</th>
<th>Graduate Students</th>
<th>Session 1</th>
<th>Session 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>E1 M1 D1</td>
<td>D5 M5 E5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>M1 D1 E1</td>
<td>E5 D5 M5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>D1 E1 M1</td>
<td>M5 E5 D5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>E1 M1 D1</td>
<td>D5 M5 E5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>M1 D1 E1</td>
<td>E5 D5 M5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>D1 E1 M1</td>
<td>M5 E5 D5</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>E2 M2 D2</td>
<td>D1 M1 E1</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>M2 D2 E2</td>
<td>E1 D1 M1</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>D2 E2 M2</td>
<td>M1 E1 D1</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>E2 M2 D2</td>
<td>D1 M1 E1</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>M2 D2 E2</td>
<td>E1 D1 M1</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>D2 E2 M2</td>
<td>M1 E1 D1</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>E3 M3 D3</td>
<td>D2 M2 E2</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>M3 D3 E3</td>
<td>E2 D2 M2</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>D3 E3 M3</td>
<td>M2 E2 D2</td>
</tr>
</tbody>
</table>

Table 3.4 Passage numbers

<table>
<thead>
<tr>
<th>Serial No</th>
<th>Passage No</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>1</td>
</tr>
<tr>
<td>E2</td>
<td>2</td>
</tr>
<tr>
<td>E3</td>
<td>3</td>
</tr>
</tbody>
</table>
3.3.2 Test procedure

At the beginning of the first session, participants were given brief instructions, and they signed the release form. Then three envelopes were given to each participant. Two of the envelopes had “For Session 1” and “For Session 2”, respectively, and a serial number on the outside. Each of the two envelopes contained three passages in accordance with the test passage order arrangement, and each passage was printed on a page by itself. The third envelope was for holding the finished evaluation sheets and translations. After the participants wrote down their names on the three envelopes, the “For Session 2” envelopes were gathered by the tester for use in the second session.

Participants were requested to take out one passage from the envelope at a time. The following procedure was followed:

1. Participant reads the first passage
2. Participant does pre-translation rating, and puts the pre-translation form into the third envelope
3. Participant translates the passage
4. Participant does post-translation rating, and places the passage, translation and rating in the third envelope

5. Participant takes out the second passage

6. Back to step 1 until the three passages are all completed

The test passages were carefully folded so that participants would not know the contents of other passages they had not opened. They were not allowed to go back into the third envelope to work on a previous translation or rating.

Participants were told beforehand to bring their own dictionaries to the test. The translation brief was that their translations would be published in the school newspaper. There was no time constraint, although the class time was 110 minutes. If there were a time constraint, participants would not pay enough attention to the questionnaires, and it would violate some statistical tests’ assumptions. For instance, internal consistency reliability, which can be for checking a rating scale’s internal consistency, is not an appropriate reliability estimate if the test is characterized as a speeded test (Chen & Krauss, 2004a, p. 502).

As the participants’ mother tongue was Mandarin Chinese, the instructions and ratings were given in Chinese. A professor in translation in Beijing was invited to make sure that the Chinese version and the English version of the instructions were equivalent.

3.3.3 Pilot study

In order to determine the viability of the pre-translation and post-translation ratings and the time needed for the experiment, a pilot study was conducted among 25
English-major juniors from a university in Harbin, China. They were requested to translate three 130-word English passages into Chinese. The students finished the three passages and ratings in 50-90 minutes. One revision after the pilot study involved the format of the rating scale of NASA Task Load Index. The original format is like the following one:

<table>
<thead>
<tr>
<th>Mental Demand</th>
<th>How mentally demanding was the task?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

Fig. 3.3 Original NASA-TLX subscale

It is a 20-point rating scale. Several participants mentioned that this unnumbered scale was not very clear and they would prefer a numbered one. Also, this author found it time-consuming to enter participants’ ratings on unnumbered scales into the computer. Therefore, the following format was used in the formal experiment:

<table>
<thead>
<tr>
<th>Mental Demand</th>
<th>How mentally demanding was the task?</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td>High</td>
</tr>
</tbody>
</table>

Fig. 3.4 Adapted NASA-TLX subscale

The TLX Own Performance (OP) subscale is as follows:

<table>
<thead>
<tr>
<th>Performance</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>Poor</td>
</tr>
</tbody>
</table>

Fig. 3.5 Original Performance subscale in NASA-TLX
In other subscales, a low number indicates a low level of mental demand, effort, or frustration. In this OP subscale, a low number is associated with good performance, because lower workload is usually accompanied by better performance. However, this can cause confusion “due to population stereotypes about whether one’s own performance should have a high number associated with good performance and a low number associated with bad performance” (Hart & Staveland, 1988, p.175). For this reason, this subscale was reversed in this study, as follows:

<table>
<thead>
<tr>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
</tr>
<tr>
<td>Good</td>
</tr>
</tbody>
</table>

Fig. 3.6 Adapted Performance subscale in NASA-TLX

Values on this subscale would be reversed back to the original format (i.e., Good-Poor) by subtracting the values from 10. For instance, if a participant assigned a 4 on the Poor-Good scale, it would be a 6 (=10-4) on the original Good-Poor scale.

### 3.3.4 Grading the translations

Participants’ translations were manually keyed verbatim into the computer by this author and several helpers, and then double-checked by this author. Compared with handwritten translations, an electronic version presents no illegibility problems and improves data quality. In the professional world, no translator translates on paper, and their handwritings have nothing to do with translation quality. More importantly, the graders would not be influenced by the participant’s handwriting. All translations were
uniformly formatted so that the format would not be a factor influencing translation assessment.

In this study, participants’ translations were evaluated by three independent graders, all of whom had translation teaching experience and had translated at least two books. The graders were requested to use both holistic and analytic methods. To facilitate grading and analysis, this author and a collaborator developed a software program called TransGrading (see http://www.transgrading.com). It allows the user to use holistic and/or analytical scoring methods; the error category list and holistic category list are customizable; it can calculate error deduction points and total scores, and export errors, error deduction points, and total scores to Microsoft Excel for further analysis.

The error category list used in this study was based on the ATA (American Translation Association) error category list with some modifications tailored for English-Chinese translation, as follows (Figure 3.7):
The graders were required to identify translation error and then assign 1, 2, 4, 8, or 16 points to each error. To help them make decisions, the flowchart for error point decisions by ATA was also adopted (see Appendix D).

The holistic score for each translation was from 0 to 10, with 0 indicating a poor translation and 10 indicating an excellent translation. Only integer scores were used. See Appendix E for the holistic scoring criteria.

As mentioned in the previous chapter, grader training and selecting benchmark responses are indispensable for improving grading accuracy. In this study, each of the 15 English source texts was translated by one of the graders into Chinese, then edited and
approved by the other two graders. A consensus was established among them regarding what an ideal translation of each text would look like. Having familiarized themselves with the holistic and analytic scoring criteria, the graders graded one translation of each of the 15 source texts. Their scores were then compared, and differences in error categories assigned and points deducted were discussed among them in order to reduce their disagreements. An agreed-upon holistic grade was assigned to each of the 15 translations. These 15 translations and 15 reference translations worked as benchmarks for the grading.

The graders graded the translations using TransGrading, whose interface is as follows (Figure 3.8):

![Fig. 3.8 The interface of TransGrading](image_url)
The graders first used the analytic method, identified a mistake and assigned a point to it. They didn’t need to add up the points deducted as the software program could do it. Then they read the whole translation quickly and gave a holistic grade.

All the translations of the same source text were put in one target file so that the grader could be more consistent when assigning grades to different translations of the same source text. In addition, the graders were requested to compare the student translations with the reference translation, sentence by sentence, especially when they were grading the first three translations. This would make them more familiar with the content, and it proved that this also helped graders spot the omitted sentences in the translation.

### 3.3.5 Calculating interrater reliability

Despite the scoring guides, grader training and benchmark responses, graders often do not assign the same grades to the same translations, partly because people have different ideas about what exactly constitutes translation quality (e.g., House, 2009). In order to test whether the graders were consistent and their ratings were generalizable to other graders, the interrater reliability among the graders needed to be determined.

Among the three most common types of consistency estimates, i.e., (1) correlation coefficients (e.g., Pearson and Spearman), (2) Cronbach's alpha, and (3) intraclass correlation, Cronbach's alpha correlation coefficient is suitable for this study (see Multon, 2010). Cronbach’s alpha correlation coefficient normally ranges from 0 for a completely unreliable test to 1 for a completely reliable test. An acceptable level for Cronbach's alpha is .70. Cronbach’s alpha for the current study was .98. Mean for individual
translations was 7.48, variance=1.36, and range=6.67. This indicates that the three graders were consistent and their ratings were highly reliable. The graders’ ratings of each translation were then averaged, and the average was taken as the translation’s quality score.

3.3.6 Checking for normality and outliers

The normal distribution plays a very important role in statistics. First, many variables in psychological measurement (e.g., measures of reading ability) are normally distributed, so the properties of the normal curve apply to the data. This simplifies the mathematical procedures for computing probabilities. Second, most statistical procedures (e.g., t-tests, analysis of variance, tests for regression coefficients) begin with the assumption of the normal distribution to make inferences about the population based on the sample. If this assumption is violated, then the probabilities associated with a normal distribution are not valid (Urdan, 2010; Yang, 2007). Therefore, we need to test whether the normal distribution holds in a study. Besides plotting methods (e.g., Q-Q plot), there are over 40 formal procedures for testing specifically for normality (see Thode, 2002). The two very common and recommended methods are skewness and kurtosis, first used by Fisher in 1930.

Outliers refer to scores or values that are unusually large or small relative to other scores or values in the data set. They may be genuine observations from participants, or they may result from human error, instrument failure, or other causes. For instance, a scale of 0-10 was used, and now there is a score of 14 or -5 in the data set; in this case, 14 and -5 are outliers. Typing mistakes are the most frequent source of errors when
transcribing and entering data into the computer (Petrie & Sabin, 2009, p. 18). It is important to detect whether there are outliers in the data set, as they may impact the calculated mean values and standard deviations. What is more important is that the presence of outliers “means that the distribution is not exactly a normal distribution;” this is “very problematic for ‘classic’ statistics (and by this I mean parametric statistical methods which do not use robust methods of estimation) because data are assumed to follow a normal distribution exactly” (Larson-Hall, 2010, p. 91).

There are different strategies for identifying outliers, including employing a standard deviation criterion, a boxplot, or one of the over 50 inferential statistical tests (see Sheskin, 2010). The most frequently used method for identifying outliers in a small or moderate-sized data set is visually checking the data set by eye (Rencher, 2002, p. 100). A less subjective way is using a statistical software program to check the range of numerical data. If a value lies outside the range then it warrants investigation.

Table 3.5 was generated by SPSS (version 20) describing the data set used in this study.

Table 3.5 Descriptive statistics for the 7 variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Statistic</td>
<td>Statistic</td>
<td>Statistic</td>
<td>Statistic</td>
<td>Statistic</td>
<td>Std. Error</td>
</tr>
<tr>
<td>predictedDifficulty</td>
<td>609</td>
<td>0</td>
<td>9.0</td>
<td>4.387</td>
<td>1.7002</td>
<td>.082</td>
<td>.099</td>
</tr>
<tr>
<td>timeSpentMinutes</td>
<td>609</td>
<td>0</td>
<td>51</td>
<td>16.40</td>
<td>5.971</td>
<td>.012</td>
<td>.099</td>
</tr>
<tr>
<td>MentalDemand</td>
<td>609</td>
<td>0</td>
<td>9.0</td>
<td>4.405</td>
<td>1.8603</td>
<td>.108</td>
<td>.099</td>
</tr>
<tr>
<td>Effort</td>
<td>609</td>
<td>1.0</td>
<td>9.0</td>
<td>4.685</td>
<td>1.8744</td>
<td>.034</td>
<td>.099</td>
</tr>
<tr>
<td>Frustration</td>
<td>609</td>
<td>0</td>
<td>9.0</td>
<td>3.530</td>
<td>1.8203</td>
<td>.201</td>
<td>.099</td>
</tr>
<tr>
<td>Performance</td>
<td>609</td>
<td>5</td>
<td>10.0</td>
<td>5.833</td>
<td>1.6182</td>
<td>-.233</td>
<td>.099</td>
</tr>
<tr>
<td>transitionGrade</td>
<td>609</td>
<td>3.00</td>
<td>10.0</td>
<td>7.5284</td>
<td>1.18453</td>
<td>.494</td>
<td>.099</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>609</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The ranges for variables predictedDifficulty, mentalDemand, Effort, Frustration, Performance, and translationGrade were all 0-10. We can see that there were no errors in the Maximum column as none of the numbers exceed 10. However, the several zeros in the Minimum column look suspicious. They were mainly caused by missing values. A “data filter” check for missing values using Microsoft Excel showed that two participants did not complete their pre-translation ratings; one of them missed 3 and the other missed 2, hence the 5 missing values in the predictedDifficulty column. There were 9 missing values in the column of timeSpentMinutes (the variable for measuring how much time they spent on the translation). There were 7 zeros in the column of Frustration (a NASA-TLX subscale), which were not missing values. It turned out that 6 of the 7 zeros were from one graduate student, and 1 from another graduate student.

This table also includes the skewness and kurtosis of each variable. The two are used to test whether a distribution is a normal distribution. For a normal distribution, their values would be close to zero, but there is an acceptable range. An oft-quoted range is by Kline (2005), who recommends that “absolute values of the kurtosis index greater than 10.0 may suggest a problem and values greater than 20.0 may indicate a more serious one” (p. 50), while absolute values of skew greater than 3.0 indicate the distribution is extremely skewed.

From the above table, we can see that most skewness and kurtosis values are within the range of -1.0 to +1.0. The only variable whose absolute skewness or kurtosis value is greater than 1.0 is timeSpentMinutes, but its kurtosis value is still smaller than 10.0. In a word, there was no violation of the normality assumption.
This chapter describes the methods used in this study, including participants, test passages, translation rating scales, experimental design, test procedure, pilot study, translation grading methods, and normality and outlier checking. The statistical analysis methods, which will be described in detail in next chapter, included frequency distributions, correlational analysis, and multiple regression analysis.
CHAPTER 4

Results

As mentioned in preceding chapters, readability formulas were tested to see whether they could be used to predict the translation difficulty level of a text, with the thought that, if they could not, a model would be developed using multiple regression analysis. However, in order to test the effectiveness of readability formulas or develop a model to predict the translation difficulty level of a text, we first needed to know how to measure translation difficulty.

The specific research questions included 1) whether NASA-TLX (Task Load Index) is reliable in measuring translation difficulty, 2) whether translation quality scores can be used to measure or represent translation difficulty, 3) whether time on task can be used to measure or represent translation difficulty, 4) whether Flesch Reading Ease formula (or readability formulas in general) can be used to predict a text's level of translation difficulty, 5) How to know a text's level of translation difficulty for a translator without having the text translated first, and 6) what the sources of translation difficulty in terms of translation errors are. Those questions will be addressed in turn in the following sections.

4.1 Checking NASA-TLX’s reliability

As scales are reliable with some groups but may be unreliable with other groups (Pallant, 2011), it is necessary to check their reliability. This is especially so in this case
for two reasons. First, to the best of this author’s knowledge, NASA-TLX has not been used to evaluate readability or translation difficulty. Second, only four of its six subscales (i.e., mental demand, physical demand, temporal demand, satisfaction in performance, effort, and frustration level) were adopted in this study.

Reliability concerns “the relative amount of random inconsistency or unsystematic fluctuation of individual responses on a measure” (Chen & Krauss, 2004b, p. 952). These unsystematic fluctuations lead to inconsistent responses by a respondent, and do not possess systematic patterns (ibid.). The three major types of reliability are interrater reliability, test-retest reliability, and internal consistency reliability. The test-retest reliability of a scale is evaluated by administering the scale to the same people on two different occasions. In this case, a participant translated the same passage only once; therefore, the test-retest reliability was not tested.

Internal consistency refers to the degree to which the items on a scale are all measuring the same underlying construct. It can be measured in a number of ways, such as Cronbach's Alpha, Kuder Richardson 20 (KR20), Split-Half Reliability, and stratified alpha. Of these, Cronbach’s Alpha (Cronbach, 1951) is the most commonly used one, and “should be used with all instruments that employ a Likert-type scale” (Lester & Bishop, 2000, p. 24). It measures the average correlation among all of the items that make up the scale, and produces values between 0 and 1, with a higher value indicating a higher degree of internal consistency. The generally accepted minimum Alpha coefficient is .70 (Nunnally, 1978). It should be noted that the value of the Alpha coefficient also depends
on the length of the scale; more items on a scale will lead to a larger Alpha coefficient (Lester & Bishop, 2000).

Recall that the two ends of the Performance subscale in NASA-TLX were reversed in data collection to avoid confusing the participants. The two ends were changed back to the original arrangement in NASA-TLX for analysis by subtracting the values from 10. For instance, if a participant assigned a 4 on the Poor-Good scale, it would be a 6 (=10-4) on the Good-Poor scale.

In this study, as 15 source texts were used and participants translated different source texts, the Alpha coefficient for each passage was calculated using SPSS. The following are the results (Table 4.1):

Table 4.1 Descriptive statistics for the 15 passages

<table>
<thead>
<tr>
<th>Passage</th>
<th>Participants/Cases</th>
<th>Alpha Coefficient</th>
<th>Mean</th>
<th>Variance</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47</td>
<td>0.73</td>
<td>18.71</td>
<td>23.18</td>
<td>4.82</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>0.86</td>
<td>14.16</td>
<td>30.50</td>
<td>5.52</td>
</tr>
<tr>
<td>3</td>
<td>43</td>
<td>0.80</td>
<td>16.30</td>
<td>24.16</td>
<td>4.91</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>0.74</td>
<td>15.97</td>
<td>17.87</td>
<td>4.23</td>
</tr>
<tr>
<td>5</td>
<td>36</td>
<td>0.82</td>
<td>15.35</td>
<td>31.11</td>
<td>5.58</td>
</tr>
<tr>
<td>6</td>
<td>47</td>
<td>0.74</td>
<td>15.34</td>
<td>24.80</td>
<td>4.98</td>
</tr>
<tr>
<td>7</td>
<td>45</td>
<td>0.80</td>
<td>14.28</td>
<td>23.27</td>
<td>4.82</td>
</tr>
<tr>
<td>8</td>
<td>43</td>
<td>0.73</td>
<td>17.70</td>
<td>24.92</td>
<td>4.99</td>
</tr>
<tr>
<td>9</td>
<td>32</td>
<td>0.73</td>
<td>15.59</td>
<td>20.39</td>
<td>4.52</td>
</tr>
<tr>
<td>10</td>
<td>36</td>
<td>0.83</td>
<td>18.76</td>
<td>33.01</td>
<td>5.75</td>
</tr>
<tr>
<td>11</td>
<td>47</td>
<td>0.72</td>
<td>21.27</td>
<td>25.57</td>
<td>5.06</td>
</tr>
<tr>
<td>12</td>
<td>45</td>
<td>0.71</td>
<td>19.97</td>
<td>21.72</td>
<td>4.66</td>
</tr>
</tbody>
</table>
The average Cronbach's Alpha coefficient was .77, and none of the Alpha coefficient was below .70. This indicates a satisfactory degree of NASA-TLX’s internal consistency. The coefficient of .77 means that the four items constituting the scale accounted for 77% true measurement of workload among the participants.

This Alpha coefficient also indicates that the scale contained about 23% (=1-77%) measurement error (i.e., unsystematic fluctuations). The measurement error might be attributed to one or more items on the scale that do not belong to the set, or a small amount of error may exist among all of the items (Lester & Bishop, 2000). In order to identify the measurement error, item analysis procedures including “Corrected Item-Total Correlation” and “Cronbach's Alpha if Item Deleted” were used. The following are the averages of the 15 passages, computed together with the above Cronbach’s alpha coefficients by SPSS (see Table 4.2).

Table 4.2 Item-total statistics for NASA-TLX’s internal consistency

<table>
<thead>
<tr>
<th>Subscale/Item</th>
<th>Corrected Item-Total Correlation</th>
<th>Cronbach's Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental Demand</td>
<td>.72</td>
<td>.65</td>
</tr>
<tr>
<td>Effort</td>
<td>.70</td>
<td>.65</td>
</tr>
<tr>
<td>Frustration</td>
<td>.68</td>
<td>.66</td>
</tr>
</tbody>
</table>
“Corrected Item-Total Correlation” refers to the Pearson correlation of each individual item with the total of the remaining items on the scale, where low values indicate weaker correlations. A common rule of thumb is that if this correlation is .40 or above, this item is at least moderately correlated with the remaining items and makes a good component of this scale (Leech, Barrett, & Morgan, 2005). In this case, the subscale Own Performance’s correlation was only .29. “Cronbach's Alpha if Item Deleted” is the estimated value of Cronbach's alpha for the whole scale if the given item were removed from the analysis. We can see that if Own Performance were deleted from this scale, the Cronbach's alpha would be .86, higher than the current average Cronbach's Alpha coefficient of .77.

However, the decision regarding whether to discard an item requires subjective judgment, and the instrument constructor can determine the criterion for including, discarding, or rewriting items to be included in the final scale (Lester & Bishop, 2000, p. 25). According to Hart & Staveland (1988), creators of NASA-TLX, Own Performance (OP) “was considered a priori as moderately related to workload,” and OP ratings were “relatively independent of other ratings, in comparison to the general finding of high statistical associations” (p. 164-165).

As a result, the total of the scores on the four subscales can be the translation difficulty score. Alternatively, the average of the scores on the four subscales can be
taken as the translation difficulty score so that the range of translation difficulty scores will be 0-10.

### 4.2 Translation difficulty level and translation quality score

In this section, the major research question is whether translation quality scores can be used to measure or represent translation difficulty.

As mentioned earlier, the participants in this study included 49 third-year undergraduate students in translation/English and 53 first-year graduate students in translation. Although it was tempting to equate graduate students with good translators, and undergraduate students with poor translators, it was necessary to look more closely into how undergraduate and graduate students performed in each of the 15 passages.

An independent-samples \( t \)-test was carried out on each passage. For each passage, on average, 20 undergraduate students and 20 graduate students did the translation. The null hypothesis was that there was no significant difference in translation quality score between undergraduate and graduate students. The alternative hypothesis was that either graduate students or undergraduate students performed better than the other group. The significance level (alpha level) was .05. The results are as shown in Table 4.3:

<table>
<thead>
<tr>
<th>Passage</th>
<th>Mean Translation Quality Score</th>
<th>Standard Deviation</th>
<th>P-value (two-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Undergraduate</td>
<td>Graduate</td>
<td>Undergraduate</td>
</tr>
<tr>
<td>1</td>
<td>7.10</td>
<td>7.18</td>
<td>.86</td>
</tr>
<tr>
<td>2</td>
<td>8.13</td>
<td>7.7</td>
<td>1.01</td>
</tr>
<tr>
<td>3</td>
<td>7.53</td>
<td>7.42</td>
<td>1.10</td>
</tr>
</tbody>
</table>
We can see that overall, undergraduate students ($M=7.63$) outperformed graduate students ($M=7.38$), and graduate students performed better than undergraduate students only in passages 1 (Easy), 6 (Medium), and 11 (Difficult). None of the $p$-values is below .05. Therefore, there was no significant difference in translation quality score between the undergraduate and graduate students.

Thus, the participants’ performance (i.e., their translation quality scores) was used to classify them into two groups: high scorer group and low scorer group. Although the two groups may comprise the top and bottom 50% of scorers, researchers in the field of test development and other fields often take the upper 27% of the scorers as the high-ability group and the lower 27% as the low-ability group (see Kelley, 1939; Feldt, 1961). This sampling method is called the extreme groups approach, and it is often used to
achieve greater statistical power in subsequent hypothesis tests (Preacher, Rucker, MacCallum, & Nicewander, 2005, p. 178). In this study and for this research question, the upper and lower 27% of the scorers on each of the 15 passage were put in the high scorer group ($M=8.81$) and low scorer group ($M=6.15$), respectively.

Figure 4.1 presents the mean self-assessed translation difficulty scores of high and low scorers in each of the 15 passages. We can see that the two lines are intertwined. This means that the translation difficulty level evaluated by good translators is not consistently lower or higher than that by poor translators.

![Figure 4.1](image)

Fig. 4.1 High scorers vs. low scorers in translation difficulty score

A correlation analysis between translation difficulty score and translation quality score was conducted. It was found that translation difficulty score was negatively and
weakly correlated to translation quality score, $r(607) = -.12, p < .01$. This indicates that the lower the translation difficulty score, the higher the translation quality score. However, this relationship was almost negligible. The value of $R^2$ was .015, and it means that translation quality score could only account for 1.5% of the variation in the translation difficulty score and 98.5% of the variation in translation difficulty score could not be explained by translation quality score. Figure 4.2 shows a scatterplot of this correlation.

Fig. 4.2 Correlation between translation quality score and translation difficulty score
To sum up, translation difficulty score and translation quality score are different, and translation performance cannot be used to measure or represent translation difficulty (or workload).

### 4.3 Translation difficulty level and time on task

Translation researchers have been wondering whether longer time spent on a translation is associated with better translation quality. Hence, there is a need to test whether time on task is correlated with translation quality score or translation difficulty level.

Recall that there were 9 missing values in the column timeSpentMinutes in the data sheet. In this analysis, those 9 cases were deleted. Two correlation analyses were conducted.

The correlation analysis between time spent on a translation and translation quality score showed that the two were not significantly related, $r(598) = -.05$, $p > .05$. This means that longer time spent on a translation was not associated with better translation quality. High performers may finish a translation faster than low performers. Table 4.4 displays some descriptive statistics.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time on translation (minutes)</td>
<td>16.64</td>
<td>5.665</td>
<td>600</td>
</tr>
<tr>
<td>Translation quality score</td>
<td>7.514</td>
<td>1.18709</td>
<td>600</td>
</tr>
</tbody>
</table>

In the second correlation analysis, it was found that time spent on a translation was significantly, but weakly, related to translation difficulty level, $r(598) = .29$, $p < .001$. 
Figure 4.3 shows a scatterplot of this correlation. We can see that as translation difficulty level increased, participants tended to spend more time on the translation. However, this relationship was weak. The value of $R^2$ was .086, and it means that time spent on a translation could only explain 8.6% of the variation in the translation difficulty score.

Fig. 4.3 Correlation between translation difficulty level and time on task

4.4 Translation difficulty level and text readability

As mentioned earlier, readability formulas might be able to be used to predict the translation difficulty level of a passage. There are many readability formulas. Table 4.5 shows the scores by 5 readability formulas for each source text. Based on their Flesch readability scores, the source texts were divided into three categories: Easy (72-80), Medium (46-53), and Difficult (21-30). The three categories were clearly separated. As it is hard to prove that Flesch can be used to measure translation difficulty, the falsification
approach was adopted. If the ranking of the three categories based on Flesch is different from the ranking based on translation difficulty score, then it can be concluded that Flesch cannot be used to measure translation difficulty.

Table 4.5 Readability scores of the 15 passages

<table>
<thead>
<tr>
<th>Passage</th>
<th>Category</th>
<th>Flesch Reading Ease</th>
<th>Flesch-Kincaid</th>
<th>Gunning Fog</th>
<th>PSK Gunning Fog</th>
<th>SMOG</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>Easy</td>
<td>72</td>
<td>7.4</td>
<td>8.7</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>02</td>
<td>Easy</td>
<td>74</td>
<td>5.8</td>
<td>8.4</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>03</td>
<td>Easy</td>
<td>76</td>
<td>5.5</td>
<td>8.1</td>
<td>4.9</td>
<td>8</td>
</tr>
<tr>
<td>04</td>
<td>Easy</td>
<td>79</td>
<td>6.6</td>
<td>9</td>
<td>5.1</td>
<td>8</td>
</tr>
<tr>
<td>05</td>
<td>Easy</td>
<td>80</td>
<td>6.3</td>
<td>7.9</td>
<td>4.8</td>
<td>6</td>
</tr>
<tr>
<td>06</td>
<td>Medium</td>
<td>46</td>
<td>11.3</td>
<td>12.5</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>07</td>
<td>Medium</td>
<td>48</td>
<td>11.7</td>
<td>15</td>
<td>6.5</td>
<td>14</td>
</tr>
<tr>
<td>08</td>
<td>Medium</td>
<td>49</td>
<td>11.2</td>
<td>15.1</td>
<td>6.6</td>
<td>13</td>
</tr>
<tr>
<td>09</td>
<td>Medium</td>
<td>51</td>
<td>9.2</td>
<td>13</td>
<td>6.1</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>Medium</td>
<td>53</td>
<td>9.1</td>
<td>10.7</td>
<td>5.6</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>Difficult</td>
<td>21</td>
<td>14.3</td>
<td>18.6</td>
<td>7.5</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>Difficult</td>
<td>28</td>
<td>13.7</td>
<td>16.1</td>
<td>6.8</td>
<td>15</td>
</tr>
<tr>
<td>13</td>
<td>Difficult</td>
<td>29</td>
<td>13.2</td>
<td>16</td>
<td>6.8</td>
<td>14</td>
</tr>
<tr>
<td>14</td>
<td>Difficult</td>
<td>29</td>
<td>12.4</td>
<td>15.8</td>
<td>6.8</td>
<td>13</td>
</tr>
<tr>
<td>15</td>
<td>Difficult</td>
<td>30</td>
<td>13.4</td>
<td>17.2</td>
<td>7.1</td>
<td>15</td>
</tr>
</tbody>
</table>

The descriptive statistics for the three passage categories are presented in Table 4.6, and the line chart shown in Figure 4.4. We can see that on average, passages in the Difficult category ($M=4.66$) were much more difficult to translate than passages in the
other categories, and passages in the Medium category ($M=4.06$) were slightly more difficult to translate than passages in the Easy category ($M=4.04$). The ranking of the three categories based on Flesch was the same as the ranking based on translation difficulty score. As there are only slight differences in the results produced by different readability formulas, and the three passage categories were clearly separated in readability level, using a different readability formula would probably produce the same ranking. Therefore, the hypothesis that a readability formula can be used to measure translation difficulty could not be falsified.

Table 4.6 Descriptive statistics for the three passage categories

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>95% Confidence Interval for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Easy</td>
<td>203</td>
<td>4.04064</td>
<td>1.311086</td>
<td>3.8592</td>
</tr>
<tr>
<td>Medium</td>
<td>203</td>
<td>4.06281</td>
<td>1.309675</td>
<td>3.88156</td>
</tr>
<tr>
<td>Difficult</td>
<td>203</td>
<td>4.65640</td>
<td>1.321945</td>
<td>4.47346</td>
</tr>
<tr>
<td>Total</td>
<td>609</td>
<td>4.25328</td>
<td>1.34277</td>
<td>4.14643</td>
</tr>
</tbody>
</table>
Fig. 4.4 Relationship between passage category and translation difficulty score

Figure 4.5 depicts how translation difficulty score varies as a function of the Flesch score. On the x-axis are the Flesch scores for the 15 passages. The readability level is from low to high. As the Flesch scores for passages 13 and 14 are both 29, and passage 14 is easier to read than passage 13 according to Flesch-Kincaid, the Flesch score for passage 14 was changed to 29.1 in order to differentiate the two. We can see that passage 11 (whose Flesch score is 21), whose readability level is the lowest, is associated with the highest translation difficulty level, while passage 5 (whose Flesch score is 80), whose readability level is the highest, has a relatively low translation difficulty score.
A correlation analysis between translation difficulty score and Flesch score was carried out. The relationship between the two was found to be negatively and weakly related, Kendall tau-b = -.141, N = 609, \( p < 0.001 \). This indicates that as the readability level increases, translation difficulty score tends to decrease. However, the correlation coefficient was very small.

### 4.5 Predicting the level of translation difficulty of a text

In the previous section, readability formulas based on average sentence length and percentage of difficult words were shown to be weakly correlated to translation difficulty level. This means that average sentence length, word difficulty, and other statistically measurable factors might be useful to measure or predict the translation difficulty level of a text.
In this study, 15 test passages were used, and their mean translation difficulty scores assessed by the participants are as follows (see Table 4.7).

Table 4.7 Mean translation difficulty scores of the 15 passages

<table>
<thead>
<tr>
<th>Passage</th>
<th>N</th>
<th>Mean Translation Difficulty Score</th>
<th>Std. Deviation</th>
<th>95% Confidence Interval for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>1</td>
<td>47</td>
<td>4.67819</td>
<td>1.19</td>
<td>4.33</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>3.53889</td>
<td>1.38</td>
<td>3.12</td>
</tr>
<tr>
<td>3</td>
<td>43</td>
<td>4.07558</td>
<td>1.23</td>
<td>3.70</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>3.99219</td>
<td>1.06</td>
<td>3.61</td>
</tr>
<tr>
<td>5</td>
<td>36</td>
<td>3.83681</td>
<td>1.39</td>
<td>3.36</td>
</tr>
<tr>
<td>6</td>
<td>47</td>
<td>3.83511</td>
<td>1.24</td>
<td>3.47</td>
</tr>
<tr>
<td>7</td>
<td>45</td>
<td>3.56944</td>
<td>1.21</td>
<td>3.21</td>
</tr>
<tr>
<td>8</td>
<td>43</td>
<td>4.42442</td>
<td>1.25</td>
<td>4.04</td>
</tr>
<tr>
<td>9</td>
<td>32</td>
<td>3.89844</td>
<td>1.13</td>
<td>3.49</td>
</tr>
<tr>
<td>10</td>
<td>36</td>
<td>4.69097</td>
<td>1.44</td>
<td>4.21</td>
</tr>
<tr>
<td>11</td>
<td>47</td>
<td>5.31649</td>
<td>1.26</td>
<td>4.95</td>
</tr>
<tr>
<td>12</td>
<td>45</td>
<td>4.99167</td>
<td>1.17</td>
<td>4.64</td>
</tr>
<tr>
<td>13</td>
<td>43</td>
<td>4.10174</td>
<td>1.15</td>
<td>3.75</td>
</tr>
<tr>
<td>14</td>
<td>32</td>
<td>4.37891</td>
<td>1.34</td>
<td>3.89</td>
</tr>
<tr>
<td>15</td>
<td>36</td>
<td>4.28472</td>
<td>1.33</td>
<td>3.84</td>
</tr>
<tr>
<td>Total</td>
<td>609</td>
<td>4.25328</td>
<td>1.34</td>
<td>4.15</td>
</tr>
</tbody>
</table>

In order to develop a model (i.e., an equation) for predicting a passage’s translation difficulty level, multiple regression analysis (see, e.g., Pedhazur, 1997) is needed. According to Green (1991), the minimum sample size for multiple regression
analyses is $50 + 8 \times \text{(number of predictors)}$. This means that if there are seven predictors (i.e., independent variables), the minimum sample size would be $50 + 8 \times 7 = 106$. In this study, only 15 test passages were used and therefore a multiple regression analysis for developing a translation difficulty model could not be carried out.

To develop such a model, a researcher needs to have a large number of participants evaluate the translation difficulty level of over 100 passages and have a mean (or agreed-upon) translation difficulty score for each passage. If the participants are asked to provide a translation difficulty score for each passage immediately after they finish the translation, few researchers would be able to afford to work on this topic.

However, if there is a way to get a translation difficulty score from each participant for each passage without translating it, then it would be much easier. The research question is that whether a participant can predict the translation difficulty level of a passage.

A backward stepwise multiple regression was carried out to investigate the best pattern of variables for predicting the level of translation difficulty of a text by each participant. The predictors (i.e., independent variables) taken into account included (1) Self-predicted level of translation difficulty (by the pre-translation rating), 2) Average sentence length, (3) Number of unique words, (4) Average number of syllables, (5) Number of complex (iii+ syllable) words, (6) Number of unique iii+ syllable words, and (7) Number of unique long words. The occurrences of these language factors in each passage were counted by Readability Studio.
As mentioned in Chapter 3, there were 5 missing values in the predictedDifficulty (i.e., Self-predicted level of translation difficulty) data. These 5 cases were deleted in this analysis.

The result showed that the only independent variable that was selected for entry into the analysis turned out to be the participants’ self-predicted translation difficulty level by pre-translation ratings. The other 6 independent variables were not included as they were not significant, independent predictors of translation difficulty level. See Appendix F for the output from Statgraphics Centurion (version 16.1).

The equation of the fitted multiple linear regression model is:

\[ T = 1.81306 + 0.549266 \times P \]

Where \( T \) refers to the actual translation difficulty score for the participant and \( P \) is the participant’s self-predicted level of translation difficulty (by the pre-translation rating).

An ANOVA test showed that there was a statistically significant relationship between translation difficulty score and self-predicted level of translation difficulty at the 95.0% confidence level. The adjusted R-squared statistic was .462. This means self-predicted level of translation difficulty explained 46.20% of the variance in each participant’s translation difficulty score. In multiple regression, the guideline for effect sizes (i.e., R-squared) provided by Cohen (1988, p. 478) is that .02, .13, and .26 are for small, moderate, and large effect sizes, respectively. Therefore, .462 indicates a large effect size. Figure 4.6 displays the scatterplot of the two variables.
Fig. 4.6 Relationship between self-predicted translation difficulty level and the actual translation difficulty level

This means that if a participant predicts the translation difficulty score of a passage to 6.5 on a scale of 0 to 10 with 0 being not difficult at all and 10 being extremely difficult, the actual translation difficulty level for her will probably be 5.38 (1.81306+0.549266 x 6.5=5.383289).

4.6 Sources of translation difficulty in terms of translation errors

Section 2.2 categorized sources of translation difficulty into two groups: translation factors (including text difficulty and translation-specific difficulty) and translator factors (including a set of translation subcompetences). Between the two groups, of course, there is unavoidable overlap. The question is which sources of
translation difficulty present more difficulty than do others. This question will be dealt with from the translation competence perspective.

As noted in section 2.3.3, one of the three categories of techniques for measuring mental workload is performance measures, and one of the two performance indicators is accuracy (or number of errors). In linguistics, competence and performance are an oft-discussed dichotomy. It is usually believed that “[c]ompetence in any sphere is identified with capacity or ability, as opposed to actual performance which may only imperfectly reflect underlying capacity” (Campbell & Wales, 1970, p. 246). Chomsky (1965) proposes that the linguist needs “to determine from the data of performance the underlying system of rules that has been mastered by the [individual] and that he puts to use in actual performance” (p. 4).

Table 4.8 contains the error categories used in the grading of translations and their corresponding translation subcompetences (including source language competence, transfer competence, and target language competence). According to Neubert (2000, p. 10), transfer competence refers to the tactics and strategies of converting source texts into target texts.

Table 4.8 Translation error categories and corresponding translation subcompetences

<table>
<thead>
<tr>
<th>Error Category</th>
<th>Abbr.</th>
<th>Subcompetence Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inaccuracy</td>
<td>IN</td>
<td>Source language competence/Transfer competence</td>
</tr>
<tr>
<td>Misunderstanding</td>
<td>MU</td>
<td>Source language competence</td>
</tr>
<tr>
<td>Lexical Choice</td>
<td>L</td>
<td>Source language competence/Transfer competence/Target language competence</td>
</tr>
<tr>
<td>Omission</td>
<td>O</td>
<td>Carelessness/Source language competence/Transfer competence</td>
</tr>
<tr>
<td>-------------------</td>
<td>---</td>
<td>-----------------------------------------------------------</td>
</tr>
<tr>
<td>Addition</td>
<td>A</td>
<td>Transfer competence</td>
</tr>
<tr>
<td>Grammar/Syntax</td>
<td>G</td>
<td>Target language competence</td>
</tr>
<tr>
<td>Non-idiomatic Usage</td>
<td>U</td>
<td>Target language competence</td>
</tr>
<tr>
<td>Typo/Incorrect character</td>
<td>T</td>
<td>Target language competence/Carelessness</td>
</tr>
<tr>
<td>Punctuation</td>
<td>P</td>
<td>Target language competence</td>
</tr>
<tr>
<td>Cohesion/Coherence</td>
<td>C</td>
<td>Target language (textual) competence</td>
</tr>
<tr>
<td>Register/Style</td>
<td>RS</td>
<td>Target language (textual) competence</td>
</tr>
<tr>
<td>Other</td>
<td>OTH</td>
<td></td>
</tr>
<tr>
<td>Bonus point</td>
<td>B</td>
<td></td>
</tr>
</tbody>
</table>

There are errors which may be caused by the translator’s carelessness, e.g., omissions and typos. They do not reflect the participants’ underlying capacity, namely, competence. Other error categories would be assumed to reflect the participants’ translation competence.

All the error segments together with their error categories and points deducted for each error recognized by the three graders in TransGrading were exported to Microsoft Excel, as shown in Figure 4.7.
It should be noted that graders may assign different error categories to the same error segment despite all the detailed scoring guides, examples, and grader training. For instance, if a translator renders “the Smiths” into the equivalent of “Mrs. Smith” in the target language, it may be taken as “Misunderstanding” by one grader and “Inaccuracy” by another. Take another example. If “baseball” was translated into “basketball”, it might be a careless typo or a “Lexical Choice” problem. In addition, one grader may recognize more problems than does another. For these reasons, there was unavoidable variation between the graders in their grading of the 609 translations. Therefore, their average values were used in the following calculation.

Figure 4.8 presents the frequency of each error category and the total points deducted under each category.
We can see that the top six error categories in terms of frequency are Inaccuracy (IN), Lexical Choice (L), Misunderstanding (MU), Non-idiomatic Usage (U), Grammar/Syntax (G), and Omission (O). The top six error categories in terms of total points deducted are Inaccuracy (IN), Misunderstanding (MU), Lexical Choice (L), Omission (O), Grammar/Syntax (G), and Non-idiomatic Usage (U). The six error categories in both groups are the same, though in slightly different ranking order.

Table 4.9 presents the error categories, their corresponding translation subcompetence categories, and total points deducted (in percentage).

Table 4.9 Translation subcompetence categories and error points deducted

<table>
<thead>
<tr>
<th>Error Category</th>
<th>Subcompetence Category</th>
<th>Points Deducted (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inaccuracy</td>
<td>Source language competence/Transfer competence</td>
<td>28.09%</td>
</tr>
<tr>
<td>Misunderstanding</td>
<td>Source language competence</td>
<td>22.11%</td>
</tr>
<tr>
<td>Lexical Choice</td>
<td>Source language competence/Transfer competence/ Target language competence</td>
<td>14.61%</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Omission</td>
<td>Carelessness/Source language competence/Transfer competence</td>
<td>8.87%</td>
</tr>
<tr>
<td>Addition</td>
<td>Transfer competence</td>
<td>3.40%</td>
</tr>
<tr>
<td>Grammar/Syntax</td>
<td>Target language competence</td>
<td>7.56%</td>
</tr>
<tr>
<td>Non-idiomatic Usage</td>
<td>Target language competence</td>
<td>6.30%</td>
</tr>
<tr>
<td>Typo/Incorrect character</td>
<td>Target language competence/Carelessness</td>
<td>3.03%</td>
</tr>
<tr>
<td>Punctuation</td>
<td>Target language competence</td>
<td>2.40%</td>
</tr>
<tr>
<td>Cohesion/Coherence</td>
<td>Target language (textual) competence</td>
<td>2.33%</td>
</tr>
<tr>
<td>Register/Style</td>
<td>Target language (textual) competence</td>
<td>0.86%</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>0.46%</td>
</tr>
</tbody>
</table>

From this table, we can infer that participants’ deficiency in target language competence accounts for over 20% of the points deducted, and their deficiency in source language competence, over 20%. Their deficiency in transfer competence also accounts for a considerable portion of the points deducted.

One of the post-translation survey questions was that: Which of the following was more difficult for you? There were three choices: 1) Source text comprehension, 2) Reverbalization in the target language, and 3) The two were equally difficult (see Appendix B). The result was showed in Figure 4.9.
That is, only 13% of the responses pointed towards “Source text comprehension” as the aspect more difficult; 77% towards “Reverbalization in the target language” as more difficult. “Reverbalization in the target language” roughly corresponds to transfer competence and target language competence. From this perspective, the participants’ deficiency in transfer competence plays a significant role in accounting for the points deducted.
CHAPTER 5

Discussion

This chapter elaborates on the results presented in the previous chapter, and offers some explanations for some of the main findings. The discussion ties together results from the six research questions into several key points. The final section of this chapter addresses some limitations of the study and offers some suggestions for further research.

5.1 Discussion of Findings

The purpose of this study was to determine the factors causing translation difficulty and to explore ways to measure the translation difficulty level of source texts (English in this case). It was guided by six research questions. Here is a summary of the main findings.

1) NASA Task Load Index (TLX), a multidimensional scale for measuring subjective workload, can be used to assess translation difficulty for the translator, and it was shown to be reliable in this study.

2) Accuracy (i.e., translation quality score in this case), which is a commonly used performance measure for deriving an index of difficulty, was found to be an unreliable indicator of the level of translation difficulty as measured by NASA-TLX.

3) Speed (i.e., time-on-task in this case), which is also a commonly used performance measure for deriving an index of difficulty, was found to be significantly, but weakly, related to the level of translation difficulty as measured by NASA-TLX.
4) The readability score of a text as measured by Flesch Reading Ease formula was found to be weakly correlated with its translation difficulty score as measured by NASA-TLX. That is, a text’s readability only partially accounts for its level of translation difficulty.

5) A formula was developed using multiple regression to predict a text’s level of translation difficulty (as measured by NASA-TLX) for a translator by using the translator’s pre-translation rating. This will greatly facilitate future studies on developing a translation difficulty formula.

6) Sources of translation difficulty were diagnosed in terms of translation errors, and it was shown that some sources of translation difficulty present more difficulty than do others.

### 5.1.1 Readability formulas and translation difficulty

As mentioned in Section 2.3.1, readability formulas typically use one to three factors, and vocabulary difficulty and sentence length are the strongest indexes of readability. Compared with sentence length, vocabulary difficulty (i.e., word familiarity or word length) plays a more important role. This is consistent with the result of one post-translation survey question, that is, which of the following aspects in source text comprehension created more problems for you? The result indicated that among the three choices (i.e., lack of background knowledge, difficult words, long sentences—see Appendix B) “difficult words” presented the greatest difficulty to the participants, followed by “long sentences.”
This study found that as the readability level increased, translation difficulty score tended to decrease. However, the association between the two variables was weak. In other words, readability formulas cannot predict well the translation difficulty level of a text. There are reasons for this.

First, readability formulas are for measuring the readability of a text, and involve only reading comprehension. By contrast, translation competence includes such subcompetences as source language competence, transfer competence, and target language competence. This study found that participants’ deficiency in target language competence accounted for over 20% of the error points deducted, deficiency in source language competence, over 20%. Some participants commented in the survey that they had no difficulty understanding the source texts, but had considerable trouble conveying the source text idea in idiomatic Chinese. It can be inferred from these that deficiency in transfer competence can cause considerable difficulty to the translator.

Second, mainstream readability formulas were all developed in the USA, and the readers are English native speakers. It is still a question whether they work well for Chinese EFL learners. For instance, English teachers teach their students formal grammar and standard vocabulary, and usually do not teach colloquial, vernacular, informal English (e.g., slang) (Zimmerman, 2010). Many Chinese EFL learners would have difficulty understanding the informal idiomatic expressions like “Here is the deal”, “Here we go again”, and they would not find help in a regular English-Chinese dictionary. Yet, these idiomatic expressions present no difficulty to native English speakers and as they use simple words, readability formulas rate them as “easy”. Of course, most materials to
be translated are in formal English. But audiovisual translation, which has been increasing in volume fast in recent years, involves oral, informal English. Obviously, more empirical research is needed to determine whether these mainstream readability formulas work for Chinese EFL learners or a new readability formula needs to be developed for them.

5.1.2 Good translators and poor translators

Participants in this study included first-year graduate students and third-year undergraduate students. People would assume that the former outperforms the latter. But the fact is that overall, undergraduate students outperformed graduate students. This unexpected result needs to be reflected on, although similar findings have been found in the fields of reading and writing. For instance, in Carrell & Connor’s (1991) study, which involved twenty-three undergraduate and ten graduate students enrolled in an English as a Second Language Program in an American university, the undergraduates significantly outscored the graduate students in reading and writing tests.

In this study, the participants were from two universities. All undergraduate students were from a university which had a very strong foreign language/translation program. Twelve graduate students were also from that university, and the rest from another university which had a moderately strong translation program. That is, students from the better program were probably academically more able than their counterparts from the other program. This may partially account for this result.

As mentioned previously, translation competence includes source language competence, transfer competence, and target language competence. In this case, target
language competence refers to participants’ competence in Chinese. Do graduate students have a better competence in Chinese than do undergraduate students? The answer is negative, as Chinese college students (especially graduate students) in foreign language programs usually do not take courses in Chinese language.

As for the fact that the graduate students had taken more courses in translation (especially in translation theories) than the undergraduate students, Pym (2010) says that “there is no empirical evidence” for the claim that “a translator who knows about different theories will work better than one who knows nothing about them”, because “[a]ll translators theorize, not just the ones who can express their theories in technical terms” (p. 4).

This result has implications for translation process research. Translation process researchers tend to take professional translators as “experts” and translation students as “novices.” For instance, PACTE group (2005) defines expert translators as “translators with at least 6 years of experience in a variety of fields for whom translation is their main professional activity and provides at least 70% of their income” (p.611). However, “translation expert” is a misnomer.

To become an expert in an area, one needs deliberate practice, i.e., engaging in tasks with goals that exceed the current level of performance. It occurs only under the following conditions, when (a) there is a well-defined task, (b) the task is of appropriate difficulty for the individual, (c) there is informative feedback, and (d) there are opportunities for repetition and the correction of errors (Ericsson 1996, from Shreve, 2002, p. 29). Process researchers have already found a phenomenon which has to do with
this expertise rule. By comparing her former TAP study (1992) and Gerloff’s TAP study (1988), Jääskeläinen (1996) concluded that (1) translation does not always get easier as professional experience increases, and (2) professional translators do not always succeed better than non-professionals. Kiraly (1995) explained this phenomenon: “As the translation student moves further along the evolutionary path from [novice] translator to high quality professional translator, the skills that are required are (a) less likely to be acquired by repeated practice, (b) less likely to develop naturally without specific training and pedagogical intervention, and c) more likely to involve translation quality at levels beyond that of mere semantic and syntactic correctness” (p.111). In real life, few professional translators get systematic feedback from the clients or proofreaders (Englund Dimitrova, 2005). Thus, the quality of their translation manifests no sign of improvement. In addition, source texts involve different text types and subject matters, and translators usually specialize in only a few domains. From this perspective, this is no “translation expert.” Translation process researchers need to distinguish good translators and poor translators by their performance rather than by their experience.

5.1.3 Workload and performance

As mentioned in Section 2.3.3.2, speed (i.e., time-on-task) and accuracy (i.e., translation quality score in this case), which are performance measures, are two commonly used workload indicators. The two will be discussed below.

In expertise studies, it is a consensus that experts can generate the best solution faster and more accurately than non-experts. As noted previously, in the field of translation, there have been mixed findings regarding whether professional translators
translate faster than non-professionals. This study found that time spent on a translation and translation quality score were not significantly related. High scorers and low scorers could not be distinguished by their time on task.

From the problem-solving perspective, the reason may be that translation involves solving ill-defined problems (Englund Dimitrova, 2005). There are well-defined problems and ill-defined problems. Well-defined problems are those problems whose goals, solution paths, and obstacles to solution are clear based on the information given while ill-defined problems are characterized by no single correct solution and lack of a clear problem statement as well (Davidson & Sternberg, 2003). Based on how problems are found, Getzels (1982, from Pretz, Naples, & Sternberg, 2003) classified problems into three types: those that are presented, those that are discovered, and those that are created. A presented problem does not need to be recognized; it just needs to be solved. A discovered problem (e.g., a pun in the source text) needs to be recognized first.

When working on the same source text, translators are not solving the same problems. For instance, Jääskeläinen (1999) found that non-professionals spent less time on the task than professionals, and she attributes this result to novices' presupposed ignorance of potential translation problems (e.g., the cultural connotation of a word). In Tirkkonen-Condit's two-subject experiment (1987), the mature student spent less time on the process but identified more problems, i.e., was more sensitized to potential problems but also more efficient in problem-solving. Texts differ in how many problems (e.g., caused by unfamiliar allusions) are implicit and need to be discovered. In L2-L1 translation, this requires that the translator have a good understanding of the source text.
In addition, as Pym’s (2003, p. 489) says, good translators excel in generating a series of more than one viable target text for a source text and selecting only one viable target text from them quickly and with justified confidence. Poor translators, by contrast, generate fewer equivalents at the above-phrase level in the target language, and engage in less decision making.

Of course, one factor that needs to be taken into account is the translation difficulty level of the source text, which is related to how many potential problems a translator needs to solve. It was found in this study that time on task was weakly related to workload (i.e., translation difficulty level). That is to say, as translation difficulty level increases, translators need to spend more time on the translation. This makes sense. But as translators have different goals, namely, expectations regarding translation quality, the time they would spend on the translation varies. This might explain why the correlation between time on task and workload was weak.

This study found that as translation difficulty level increased, the translation quality score tended to decrease. Unlike the hypothetical non-linear relationship between workload and performance depicted in Figure 2.2, the relationship between translation difficulty level and translation quality score was a linear, albeit weak, one. The hypothetical non-linear relationship involves three task categories: easy, moderately difficult, and very difficult. For easy and very difficult tasks, increased workload is not accompanied by variations in performance; for moderately difficult tasks, the relationship between the two variables is a linear one. In this study, the mean translation difficulty score was 4.25 (out of 10) with a standard deviation of 1.34, and the mean translation
quality score was 7.51 (out of 10) with a standard deviation of 1.18. Overall, the translation tasks belonged to the “moderately difficult” category, and as a result, the finding was consistent with the hypothetical relationship between workload and performance. Further research is needed to test the validity of this hypothetical relationship by including translation tasks that are more discriminating between good translators and poor translators.

5.1.4 Translation competence

As reviewed in Section 2.2.2, there are a number of typologies regarding the components of translation competence. They are more alike than different. This study found that participants’ deficiency in target language competence might account for over 20% of the error points deducted; source language competence, over 20%; transfer competence, probably over 30%. As the text type of the test passages used in this study (i.e., informational and expository) is the dominant one in the professional world, these numbers might have a moderate to high degree of generalizability although they cannot be generalized to all situations.

Different types of translation tasks draw on some components of translation competence more heavily than on others. For instance, if a text contains many allusions (e.g., to Bible, Greek mythology), this would require the translator to have relevant background knowledge to understand the associations and connotations (see Leppihalme, 1997). Otherwise, the translator would accept the allusions at face value. If a specialized text (e.g., on electronics) contains a great many terms, this would draw heavily on the
translator’s instrumental competence concerning the use of search engines, dictionaries, and software tools like SDL MultiTerm.

5.2  Limitations of this study and suggestions for further study

Although this study has added significantly to our understanding of translation difficulty, a number of unanswered questions still remain.

This study used fifteen test passages with an average length of 125 words. Although most readability formulas (e.g., Flesch, Fog, SMOG) work with passages shorter than 300 words (Bastable, 2008, p. 258), it has been recommended (e.g., by Readability Studio) that passages longer than 300 words be used so that the selections would be more representative of the texts. Of course, using longer test passage would come with the disadvantage of a smaller sample size of texts.

Text type is a very important factor in translation difficulty. In this study, it was a control variable, and only informational and expository texts were adopted. It would be interesting to investigate how text type influences the level of translation difficulty. For instance, as noted previously, translating literary and non-literary texts might involve solving different kinds of problems. There might be more one-to-one equivalence situations in non-literary translation than in literary translation, whereas in literary translation translators might be more engaged in selecting one optimal solution from several renditions than in non-literary translation. Descriptive and persuasive texts might also have differences in terms of translation difficulty. It has been found that descriptive texts are easier to read or write than persuasive texts (Carrell & Connor, 1991).
On a related note, when selecting test passages, researchers should avoid those that happen to belong to a domain that some participants routinely work in. Generally speaking, if a novice/poor translator thinks passage A is more difficult to translate than passage B, a professional/good translator will probably agree. This is the case for generalist translators (in contrast to specialized translators). For professional translators who routinely work in a domain (e.g., legal translation), things might be different. For instance, a professional generalist translator thinks passage A (a press release from a bank) is easier to translate than passage B (a legal contract). A translator who specializes in legal translation might think otherwise. This would complicate the results.

In this study, participants were asked to translate on paper. In the professional world, almost all translators work on a computer. The testing mode (i.e., paper-based versus computer-based assessment) might have an effect on their performance and translation difficulty evaluation. To date, there have been many studies examining the comparability of scores by paper-based and computer-based tests in other fields. For instance, Higgins, Russell, & Hoffmann (2005) found no significant differences in reading comprehension scores across testing modes. One difference between paper-based versus computer-based translation would be that translators working on a computer may have access to the Internet and electronic dictionaries. For instance, Powerword (by Kingsoft), a widely used electronic dictionary software program in China, contains over 150 authorized English-Chinese dictionaries (including several Oxford dictionaries). Looking up words would be much easier. In addition, translating on a computer will reduce considerably the numbers of typos and punctuation errors. For instance, half of the
punctuation errors made by the participants in this study were related to period (i.e., full stop). The period symbol in English is a dot, while in Chinese, it is a circle (。). This mistake will rarely be made if the translator works on a computer, because the English period symbol is not a punctuation mark in Chinese.

Having participants translate on computers also facilitate recording the translation process with screen recording or keystroke logging software tools. Besides translation scores and participants’ self-evaluation of translation difficulty, translation process data would be available for analysis. The researcher can locate places that have caused problems to the participants. Together with the use of TAP, it might also help graders determine whether or not an error reflects a deficiency in source language competence, transfer competence, or target language competence. It should be noted that translation problems do not necessarily manifest as translation errors.

NASA-TLX has two versions: the paper-and-pencil version and computer version. The former was used in this study. The computer version contains the original six subscales, and allows the user to assign a weight to each subscale through 15 possible pairwise comparisons of the subscales. As noted earlier, there are studies indicating that assigning weights is unnecessary. Still, it would be useful to verify this conclusion for translation difficulty research.

In order to develop a translation difficulty formula, the text features that can best predict the translation difficulty level of a text need to be combined. There is a need to recognize these text features. For instance, passive sentences might create more problems in English-Chinese translation than do active sentences. In translation process research,
researchers usually have participants translate test passages, and they select test passages based on passage length, text type, and subject matter. It is hard to find test passages that contain a great many specific translation problems (or text features) that the researcher might be interested in. For this kind of research questions, a group of sentences rather than whole passages may be more suitable.

Because of limited sample size, this study did not produce a translation difficulty formula. However, a formula was produced that will contribute to our ability to derive a translation difficulty score from each participant for each test passage without translating it. This will greatly facilitate future researchers to have a large number of participants evaluate the translation difficulty level of over 100 passages with a view to developing such a translation difficulty formula.

Research in translation difficulty can contribute greatly to our understanding of translation process in terms of relationships between text characteristics, translator behaviors, and translation quality. Finding a way to measure translation difficulty will help translation teachers prepare properly leveled passages for translation exercises, and language service providers have a better idea of the translation difficulty level of the materials.
APPENDIX A

Test Passages

Passage 1

How to Avoid Germs in Public

- Be a role model. Teach good hygiene practices in home and in school. Kids learn fast and will mimic hand hygiene habits, especially if a parent or teacher shows them the "why" and "how" they can successfully keep pesky germs at bay.
- Wash your hands for the amount of time it takes to sing "Happy Birthday" twice.
- To open the public bathroom door from the inside, grab a piece of paper towel, then dispose of it properly.
- Instead of using your hand to flush a public toilet, use your foot or a dispensable hand towel.
- After visiting a public event where you shook a lot of hands, wash thoroughly before you leave the event.
- Wash your fresh produce as if your family's lives depended upon it.

Passage 2

Breathing exercises are an ideal way to relieve stress in that they’re fast, simple, free, and can be performed by just about anyone. They can also be done anywhere and at virtually any time. These factors make stress relief breathing exercises one of my most popular and convenient tension tamers. Here’s how basic controlled breathing works:

1. Sit or stand in a relaxed position.
2. Slowly inhale through your nose, counting to five in your head.
3. Let the air out from your mouth, counting to eight in your head as it leaves your lungs. Repeat several times. That’s it!

As you breathe, let your abdomen expand outward, rather than raising your shoulders. This helps your lungs fill themselves more fully with fresh air, releasing more “old” air.

Passage 3

How to Teach a Child Table Manners
• Do as you expect the child to do. Teach by example.
• Give credit where it's due. If the child is doing a great job, tell him so. Avoid overpraising, since this can disrupt the meal and make the child self-conscious from too much attention.
• Be consistent. Make sure that what is and isn't allowed is the same at every meal. If elbows aren't allowed on the table at dinner, it should be the same at lunch.
• Teach him the basics. Before any expectations are placed on the child, inform him of the most essential table manners. Which utensil to use, what is appropriate behavior at the table, as well as anything else that is relevant can be relayed prior to having a meal.

Passage 4

How to Sneeze Without Spreading Germs
If you feel a sneeze coming, get ready. Using your hand is better than letting it go out into the air but not the best way to keep from spreading your germs. Once the sneeze or cough is on your hand, then what? Chances are you end up touching surfaces of things that other people will then come along and touch, or you may touch another person directly.

Use your shoulder instead when you have to sneeze or cough. Bend your shoulder and turn your face toward your shoulder to sneeze, covering your mouth completely but not blocking the air flowing out completely. Try the crook of your elbow as an alternate way to sneeze or cough.

Passage 5

Making mistakes during road races is very common. One of the biggest rookie mistakes in racing is going out too fast in the beginning of the race. Here are some ways that you can avoid going out too fast:
• Deliberately run your first mile slower than you plan to run the final one. Keep in mind that for every second you go out too fast in the first half of your race, you could lose as much as double that amount of time in the second half of your race.
• Don't start yourself with faster runners because you'll most likely try to keep up with them.
• Start your race at a comfortable pace and make sure you check your watch at the first mile marker. If you're ahead of your anticipated pace, slow down.

Passage 6

Why Men Snore
Snoring is noisy breathing that occurs during sleep. Women may snore too, but men snore more than women: four out of every ten men snore compared to roughly three out of ten women.

Snoring has many potential causes, including obesity, anatomical variations, and several illnesses affecting the upper airways. It results from the vibration of soft tissues of the upper airways and often is associated with obstruction of airflow.

Snoring is a problem because it affects everyone in earshot, including the person who is snoring.

Self-help ideas to prevent or minimize snoring

- If you're overweight try to lose some of it
- Stop smoking
- Try raising your head by putting pillows under it
- Sleep on your side
- Blow your nose before going to bed.

Passage 7

Weight control

Many physicians believe that obesity is one of the greatest threats to good health. Overweight people often have shorter lives and are also more apt to contract such diseases as diabetes, hypertension (high blood pressure), heart disease, arthritis, and various disorders of the digestive system. In the late 1980s scientists discovered strong evidence that there is a genetic basis for obesity—that is, that children of obese parents will probably become obese themselves. Nevertheless, it is still true that a person's weight is directly related to the amount of food eaten and to the level of energy expended. All foods are fattening if eaten to excess.

In reducing, the maximum safe body weight loss is about two pounds per week.

Passage 8

Research shows that having ambivalent friendships in your life—relationships where interactions are sometimes supportive and positive and sometimes hostile or negative—can actually cause more stress than relationships that are consistently negative! Additionally, relationship conflict and stress have been shown to have a clear negative impact on health, affecting blood pressure, contributing to heart disease, and correlating with other conditions. That’s why it’s in your best interest to minimize or eliminate negative relationships in your life.

Make a list of friendships in your life. Include everyone you think of when you think of your ‘friends’. Circle the names of people who support you when you’re down and
genuinely share your joy when good things happen to you. Now put more of a focus into the relationships you have with these people.

Passage 9

How to Get a Good Memory

• Work out regularly. Working out raises the oxygen levels in the brain, which can decrease the dangers of developing illnesses that lead to loss of memory, including heart disease and diabetes.
• Consume a nutritious diet. Follow a healthy diet that is beneficial for both the brain and the heart, with plenty of whole grains and fresh fruits and vegetables. Drink a lot of water and restrict your alcohol intake.
• Get enough sleep. Sleep deprivation prevents the brain from operating at its full potential and capacity.
• Concentrate. Good concentration, focus and attention are all vital for enhancing the memory. To allow any information to become part of your long-term and permanent memory, it is important to pay active and close attention to it, without distractions.

Passage 10

Sunburn

Sunburn results when the amount of exposure to the sun or other ultraviolet light source exceeds the ability of the body's protective pigment, melanin, to protect the skin.

Sunburn is better prevented than treated. Effective sunscreens are available in a wide variety of strengths. Most doctors recommend a sunscreen SPF (Sun Protection Factor) level of 30 or greater.

If you do get a sunburn:
• Try taking a cool shower or bath or placing wet, cold wash rags on the burn.
• Avoid products that contain benzocaine, lidocaine, or petroleum (like Vaseline).
• If blisters are present, dry bandages may help prevent infection.
• If your skin is not blistering, moisturizing cream may be applied to relieve discomfort.
• Over the counter medications, like ibuprofen, may help to relieve pain from sunburn.

Passage 11

An ecosystem is defined by organic and inorganic products that maintain the homeostasis of the biosphere. The biosphere is all organic life on the planet, including plants, animals and humans. Well-sustained ecosystems help the biosphere balance the demands of
organic life, such as consumption and respiration. An impact toward any ecosystem disrupts all aspects of an ecosystem, including the local food chain or the consumption and reproduction rate of organic life forms.

Pollution, especially man-made pollution, is one of the biggest impacts on ecosystems. Pollution is preventable, as humans actively consume and utilize materials that create waste and pollutants. These pollutants disrupt many ecosystems from their homeostasis. For example, pollution such as carbon emissions may trap greenhouse gases, heating an ecosystem beyond its typical temperature.

Passage 12

Natural Cures for Tiredness

If you have been feeling tired, fatigued or stressed, natural remedies can provide you with energy and endurance. Natural health remedies include nutritional supplements such as vitamins, minerals and plant extracts.

B-complex
The B-complex vitamin encompasses the whole spectrum of B vitamins that work together to support the nervous system, increase energy levels and protect cells from free radical damage.

CoQ10
CoQ10 is an antioxidant enzyme that is vital for manufacturing cellular energy and is present in every tissue of the body. Although the body produces CoQ10, taking this enzyme as a supplement can provide added benefits.

Ginseng
Ginseng root is known worldwide for its restorative properties. A rich source of natural chemicals, ginseng root helps combat stress and increases stamina.

Passage 13

What Determines Health

The habits and behaviors established within a family group can have a great effect on an individual's health. Good health habits are usually begun in the family. A child who is brought up in a family that practices healthful behaviors is more likely to continue the regimen as an adult.

Health is continually subjected to such internal and external challenges as varying external temperatures, bacteria, viruses, and stress. The various systems within the body must constantly adjust to these changing conditions.

Some health-influencing factors can be controlled, such as eating habits and environment. Others are not controllable, such as age and heredity. Among the many factors that can affect the health status of the individual are physical activity, nutrition, safety, smoking behavior, stress adaptation, and personal health practices.
Passage 14

Superfoods
Definition: Foods that appear to offer additional health benefits beyond simple nutrition. There really isn’t one standard definition for superfoods, but they usually are some combination of the following:

- Low in calories.
- Contain substantial amounts of omega-3 fatty acids or monounsaturated fatty acids.
- High in fiber.
- Contain phytochemicals that have been found to have the potential to prevent disease in laboratory or clinical research.
- Rich in vitamins and minerals.
- Low in unhealthy substances such as saturated fats, trans-fats or refined sugars.

For example, salmon can be considered a superfood because it contains large amounts of omega-3 fatty acids and is low in saturated fat and calories. Many fruits and vegetables are considered superfoods because they are high in fiber and phytochemicals and usually low in calories.

Passage 15

Disruptions to Ecosystems
An ecosystem refers to the interacting abiotic and biotic components of a given area. Ecosystems face disruption from natural processes, such as fires and flooding, in addition to human impacts.

Fire
Fire can be naturally occurring or artificially induced. Fires can cause disruption to ecosystems by destroying habitat and killing species, such as the case with large-scale forest fires. On the other hand, in some ecosystems wildfires help determine community structure and composition.

Flooding
In a similar way to fires, floods can cause major disruptions to ecosystems, especially large floods with high debris contents that can inundate and destroy large areas of habitat. Often, flooding occurs on a cyclical timescale across the year, allowing ecosystems time to recover.
APPENDIX B

The Post-translation Questionnaire

1. How much time did you spend on this translation?
   ____ minutes (including revision time)

2. Which of the following was more difficult for you? Please tick the appropriate box.
   □ Source text comprehension
   □ Reversalization in the target language
   □ The two were equally difficult.

3. Which of the following aspects in source text comprehension created more problems for you? Please rank the items using 1, 2, and 3, with 1 being the most difficult.
   □ Lack of background knowledge
   □ Difficult words
   □ Long sentences
   □ Others, please specify

4. Which of the following aspects in reversalization created more problems for you? Please rank the items using 1, 2, and 3, with 1 being the most difficult.
   □ Word level
   □ Sentence level
   □ Style
   □ Others, please specify
APPENDIX C

The Adapted NASA Task Load Index for Measuring Translation Difficulty

1. **Mental Demand**
   How mentally demanding was the task?
   
<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. **Effort**
   How hard did you have to work to accomplish your level of performance?
   
<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. **Frustration**
   How insecure, discouraged, irritated, stressed, and annoyed were you?
   
<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. **Performance**
   How successful were you in accomplishing what you were asked to do?
   
<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td>Good</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX D

ATA Flowchart for Error Point Decisions (Version 2009)
## APPENDIX E

### Holistic Translation Scoring Criteria

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-10</td>
<td><strong>Excellent Translation</strong></td>
</tr>
<tr>
<td></td>
<td>The translation faithfully reflects all the original passage with only 1 or 2</td>
</tr>
<tr>
<td></td>
<td>minor lexical errors. It adequately reflects the style and tone of the original</td>
</tr>
<tr>
<td></td>
<td>passage. The translation is elegant (appropriate choice of words, a variety</td>
</tr>
<tr>
<td></td>
<td>in sentence patterns).</td>
</tr>
<tr>
<td>7-8</td>
<td><strong>Good Translation with Few Inaccuracies</strong></td>
</tr>
<tr>
<td></td>
<td>The translation reflects almost all the original passage with relatively few</td>
</tr>
<tr>
<td></td>
<td>significant errors in comprehending individual words, phrases, sentences or</td>
</tr>
<tr>
<td></td>
<td>ideas. The translation is readable (generally clear, smooth and cohesive).</td>
</tr>
<tr>
<td>5-6</td>
<td><strong>Passable Translation with Some Inaccuracies</strong></td>
</tr>
<tr>
<td></td>
<td>The translation adequately reflects most of the original passage with</td>
</tr>
<tr>
<td></td>
<td>occasional errors in comprehending individual words, phrases, sentences or</td>
</tr>
<tr>
<td></td>
<td>ideas. The translation is, for the most part, readable.</td>
</tr>
<tr>
<td>3-4</td>
<td><strong>Inadequate Translation with Frequent Inaccuracies</strong></td>
</tr>
<tr>
<td></td>
<td>The translation only reflects about half of the original passage with</td>
</tr>
<tr>
<td></td>
<td>frequent errors in comprehending individual words, phrases, sentences or</td>
</tr>
<tr>
<td></td>
<td>ideas. The translation is, in some parts, unreadable.</td>
</tr>
<tr>
<td>0-2</td>
<td><strong>Poor Translation</strong></td>
</tr>
<tr>
<td></td>
<td>The translation reflects less than half of the original passage. Almost all</td>
</tr>
<tr>
<td></td>
<td>sentences contain errors in comprehending individual words. Phrases,</td>
</tr>
<tr>
<td></td>
<td>sentences or ideas. The translation is, for most part, unreadable.</td>
</tr>
</tbody>
</table>
APPENDIX F

Statgraphics Output for the Stepwise Multiple Regression Analysis

Multiple Regression - Translation difficulty score
Dependent variable: Translation difficulty score
Independent variables:
  selfPredictedDifficulty
  averageSentenceLength
  NoOfUniqueWords
  NoOfComplexWords
  AverageNumberOfSyllables
  NumberOfUniqueSyllableWords
  NumberOfUniqueLongWords

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Standard</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter Estimate</td>
<td>Error</td>
<td>Statistic</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>1.81306</td>
<td>0.114179</td>
</tr>
<tr>
<td>selfPredictedDifficulty</td>
<td>0.549266</td>
<td>0.0241138</td>
</tr>
</tbody>
</table>

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Df</th>
<th>Mean Square</th>
<th>F-Ratio</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>504.592</td>
<td>1</td>
<td>504.592</td>
<td>518.84</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>585.468</td>
<td>602</td>
<td>0.972538</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (Corr.)</td>
<td>1090.06</td>
<td>603</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-squared = 46.2903 percent
R-squared (adjusted for d.f.) = 46.2011 percent
Standard Error of Est. = 0.986173
Mean absolute error = 0.76493
Durbin-Watson statistic = 1.55623 (P=0.0000)
Lag 1 residual autocorrelation = 0.221636

Stepwise regression
Method: backward selection
P-to-enter: 0.05
P-to-remove: 0.05
The output shows the results of fitting a multiple linear regression model to describe the relationship between Translation difficulty score and 7 independent variables. The equation of the fitted model is

\[
\text{Translation difficulty score} = 1.81306 + 0.549266 \times \text{selfPredictedDifficulty}
\]
Since the P-value in the ANOVA table is less than 0.05, there is a statistically significant relationship between the variables at the 95.0% confidence level.

The R-Squared statistic indicates that the model as fitted explains 46.2903% of the variability in Translation difficulty score. The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 46.2011%. The standard error of the estimate shows the standard deviation of the residuals to be 0.986173. This value can be used to construct prediction limits for new observations by selecting the Reports option from the text menu. The mean absolute error (MAE) of 0.76493 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is less than 0.05, there is an indication of possible serial correlation at the 95.0% confidence level. Plot the residuals versus row order to see if there is any pattern that can be seen.

In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.0000, belonging to selfPredictedDifficulty. Since the P-value is less than 0.05, that term is statistically significant at the 95.0% confidence level. Consequently, you probably don't want to remove any variables from the model.
REFERENCES


*Psychometrika, 16*(3), 297-334.


process research: integrative approaches in translation studies (pp. 37-55). Amsterdam; Philadelphia: John Benjamins.


(Eds.), *Translation and cognition* (pp. 63-84). Amsterdam; Philadelphia: John Benjamins.


