Chapter 7
Measuring Difficulty in Translation and Post-editing: A Review

Sanjun Sun

7.1 Introduction

In the last decade, contributions to cognitive and psycholinguistic approaches to translation and interpreting processes have been constantly increasing. Muñoz's (2014) review of advances in this field focuses on seven, albeit overlapping, topics or research areas: competence and expertise, mental load and linguistic complexity, advances in research methods, writing, revision and metacognition, recontextualized research, and cognition beyond conscious, rational thought. Of these topics, mental load, according to Muñoz (2012), is "a construct of paramount importance" (p. 172) for translation process research, and may help us unravel the complex relationships between consciousness, problem solving, automation, and expertise; it may also establish a bridge between translation and interpreting research. It might be an overstatement to say that mental load stays at the center of this integrated view of translation process research. Nonetheless, it deserves attention and emphasis.

This article first clarifies conceptual issues and reviews difficulty, mental workload, cognitive load and other related terms, their histories and theories. Under the umbrella of cognitive science, it then reviews two lines of research, i.e., difficulty in human translation and in post-editing (PE) of machine translation. Studies concerning methods for measuring difficulty are presented and critically examined. As the author has already discussed methods for measuring difficulty in human translation elsewhere (see Sun 2015), the focus of this review is on measurement of cognitive effort in post-editing. Two assumptions in translation difficulty research are described towards the end of this article.

S. Sun (✉)
School of English and International Studies, Beijing Foreign Studies University, Beijing, China

D. Li et al. (eds.), *Researching Cognitive Processes of Translation*, New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-13-1984-6_7

7.2 Difficulty and Related Terms and Disciplines

Translation process research has been advancing through interdisciplinary research. The related disciplines include, among others, cognitive science, psycholinguistics, psychology (e.g., developmental psychology, educational psychology, assessment psychology), and neuroscience. Translation difficulty also requires interdisciplinary study. According to Newell (2001, p. 15), interdisciplinary research involves determining relevant disciplines (interdisciplines, schools of thought) by looking into each discipline and see if there is already a literature on that topic, developing working command of relevant concepts, theories and methods of each discipline, generating disciplinary insights into the problem, and integrating those insights through the construction of a more comprehensive perspective.

In the search for relevant disciplines and areas, terms play an important role. In a language, terminological variation is a common phenomenon, and the causes for the variation can be related to different origins of authors, different communicative registers, different stylistic and expressive needs of authors, contact between languages, or different conceptualizations and motivations (Freixa 2006). The variation in terminology poses a challenge for finding pertinent literature in other disciplines or sub-areas and for exploring comparability of studies. This is especially the case for research on difficulty in translation and post-editing.

As mentioned in Sun (2015), difficulty, from the cognitive perspective, refers to the amount of cognitive effort required to solve a problem, and translation difficulty can be viewed as the extent to which cognitive resources are consumed by a translation task for a translator to meet objective and subjective performance criteria. Terms similar to or synonymous with difficulty include mental load, mental workload, cognitive workload, workload, cognitive load, cognitive effort, mental effort and so forth. In psychology, the word "difficulty" often appears in the phrase "task difficulty".

Since the 1920s, psychologists (e.g., Redfield 1922) started to pay attention to workload and difficulty. Notably, Thorndike et al. (1927) focused on the measurement of difficulty, and discussed various methods in their book; Woodrow (1936) compared two scaling methods for measuring difficulty; Freeman and Giese (1940) studied whether task difficulty could be measured by palmar skin resistance. In the 1950s, information-processing models became established in psychology through Miller's (1956) article "The magical number seven, plus or minus two", which suggested that the human perceptual systems are information channels with built-in limits, and Broadbent's (1958) filter model of selective attention, which proposed that humans process information with limited capacity and an attentional filter screens out information to prevent the information-processing system from becoming overloaded (see Bermúdez 2014).

The information processing approach has had profound influence on psychology and cognitive sciences and also in task difficulty and workload research. According to Bermúdez (2014), the basic assumption shared by cognitive sciences is that "cognition is information processing" (p. 130) and "the mind is an information-

processing system" (p. xxix). Evolution of workload theory has been driven largely through empirical work conducted from a human information processing approach, which takes into account all processes studied within cognitive psychology, such as perception, attention, memory, decision making, and problem solving (Embrey et al. 2006, p. 49). In a meta-analytic review, Block et al. (2010) define cognitive load as "the amount of information-processing (especially attentional or working memory) demands during a specified time period; that is, the amount of mental effort demanded by a primary task" (p. 331).

Cognitive scientists have in recent decades suggested extending and moving beyond the basic assumption of the information process approach that the mind is an information-processing system, and have proposed dynamical systems theory and situated/embodied cognition theory. As argued by Bermúdez (2014, p. 421), these theories are "best seen as a reaction against some of the classic tenets of cognitive science" and hardly give us grounds for abandoning the whole idea of information processing.

The following paragraphs discuss mental workload, mental effort, cognitive load, mental load, and task difficulty one by one, and focus on their origin, domain, and related research.

7.2.1 Mental Workload and Mental Effort

Mental workload has been an important concept in human factors and industrial psychology. It first appeared in the 1960s (e.g., Kalsbeek and Sykes 1967), and permeated the literature in the 1970s (Vidulich and Tsang 2012, p. 243). In 1980, Wierwille and Williges prepared a report entitled "An annotated bibliography on operator mental workload assessment" for U.S. Naval Air Test Center, which included over 600 references. Mental workload assessment has been for purposes of increasing safety, reducing errors, and improving system performance (Karwowski 2012); it usually concerns high risk tasks. For example, there have been many studies on measuring the mental workload of aircraft pilots and car drivers.

A research focus in this field has been on how to measure mental workload. Wierwille and Williges (1980) in their bibliography identified 28 specific techniques in four major categories: subjective opinion, spare mental capacity, primary task, and physiological measures. Subjective measures have been very frequently used, and the most commonly used subjective measure is the rating scale. A frequently employed rating scale is NASA-TLX (Task Load Index) developed by Hart and Staveland (1988), which is the most cited work in the field of mental workload.

NASA-TLX includes six workload-related subscales, as follows: Mental Demand, Physical Demand, Temporal Demand, Effort, Performance, and Frustration Level. The Effort subscale measures "How hard [you had] to work (mentally and physically) to accomplish your level of performance?" In some literature, this is referred to as "mental effort", which means "the amount of capacity or resources that

is actually allocated to accommodate the task demands" (Paas and Van Merriënboer 1994a, p. 122). According to Paas and Van Merriënboer (ibid.), mental effort can be used as an index of cognitive load, as "the intensity of effort being expended by students is often considered to constitute the essence of cognitive load" (p. 122). This may be the case under certain circumstances, for example, when people are highly motivated. In interpreting studies, this reminds one of the Effort Models' "tightrope hypothesis" proposed by Gile (1999, 2009), according to which interpreters tend to work close to processing capacity saturation; thus, cognitive effort in interpreting often equals cognitive load. It is worth mentioning that Frustration in NASA-TLX measures an affective component of mental workload.

7.2.2 Cognitive Load and Mental Load

The term "cognitive load" has been used in psychology since the 1960s (e.g., Bradshaw 1968). It has been mainly associated with cognitive load theory (CLT) in educational research since Sweller (1988) first developed the theory. The fundamental idea behind CLT, which has become an influential instructional design theory, is that "instructional design decisions should be informed by the architecture of the human cognitive system" (Brünken et al. 2010, p. 253), and studies along this line aim at deriving empirically based guidelines for instructional design.

In the field of CLT, with regard to its measurement, cognitive load has been conceptualized in three dimensions: mental load, mental effort, and performance (Paas and Van Merriënboer 1994b). Of the three assessment dimensions, mental load refers to "the burden placed on learners due to instructional parameters" (Lusk and Atkinson 2007, p. 751). It is equivalent to Mental Demand in NASA-TLX, which measures "How much mental and perceptual activity was required (e.g., thinking, deciding, remembering, searching, etc.). The term "mental load" first appeared in the 1920s (e.g., Redfield 1922), and has been used interchangeably with mental workload in the field of psychology (e.g., Gopher 1994). There may exist some subtle difference in meaning between the two for researchers in human factors. In the term "mental load" there is an overtone of physical effort, whereas "mental workload" emphasizes the human information processing rate and the difficulty experienced (Moray 1977, p. 13).

By drawing on work in human factors, researchers in CLT often divide methods for cognitive load measurement into two groups: analytical and empirical methods (Paas et al. 2008, p. 15). Empirical methods are subjective, performance and physiological measures, while analytical methods include expert opinions, mathematical modeling, and task analysis. A well-known and extensively used subjective rating scale is a 9-point Likert scale first used by (Paas 1992), which ranges from "very, very low mental effort" (1) to "very, very high mental effort" (9).

7.2.3 *Task Difficulty*

Compared with mental workload and cognitive load, difficulty is a common term, and thus "task difficulty" has been used more frequently in various fields since the 1920s (e.g., Thorndike et al. 1927). It has been defined along two lines: (1) task difficulty refers to "the degree of cognitive load, or mental effort, required to identify a problem solution" (Gallupe et al. 1988, p. 280); (2) task difficulty is informational complexity or task complexity, and is independent of use (Rost 2006, p. 49), as in "the effects of task difficulty on mental workload". Although one may distinguish the two senses by using "subjective difficulty" and "objective difficulty" (DeKeyser 2003, p. 332), it is better to treat "task complexity" and "task difficulty" as different terms (Kuiken and Vedder 2007, p. 120).

Difficulty has been addressed in reading and writing (see e.g., Muñoz 2012) as well as in translation. For example, Wilss (1982, p. 161) distinguished between four types of translation difficulty (TD) from a pedagogical perspective:

> 1) transfer-specific TD, covering the two directions native tongue – foreign language and vice versa, 2) translator-specific TD, distinguishing two levels, one for beginners and one for advanced translators, 3) text-type-specific TD, covering at least the three particularly transfer-relevant areas, LSP translation, literary translation and Bible translation, 4) single-text-specific TD motivated by the semantically and/or stylistically complicated manner of expression of the SL author.

Nord (2005, p. 168) made similar distinctions: text-specific difficulties, translator-dependent difficulties, pragmatic difficulties, and technical difficulties. As a term, difficulty is listed in *Key Terms in Translation Studies* (Palumbo 2009).

As these terms discussed above are embedded in their respective literature, they are used interchangeably in this review.

7.3 Cognitive Science and Translation Difficulty Research

Cognitive science is a cross-disciplinary enterprise devoted to understanding mind and intelligence from an information processing perspective. It is concerned with how information is perceived, represented, transformed, stored, and communicated. Cognitive science emerged in the 1970s and draws on a host of disciplines such as philosophy, psychology, neuroscience, artificial intelligence, linguistics, and anthropology (e.g., Frankish and Ramsey 2012). It covers memory, attention, consciousness, reasoning, problem solving, decision making, metacognition, expertise, computational linguistics, cognitive ergonomics, human-computer interaction, machine translation, and so forth (see Wilson and Keil 1999). Among these research fields, cognitive ergonomics overlaps with related disciplines such as human factors, applied psychology, and human-computer interaction (Cara 1999), and according to International Ergonomics Association (2015), its relevant topics include, among others, mental workload.

Translation process research has incorporated concepts, theories and methods (e.g., metacognition, expertise studies) from cognitive sciences (see Alves 2015), and there is a need to further integrate translation process research and the cognitive sciences (Shreve and Angelone 2010), and critically examine our traditional perspectives on translation processes in terms of the frameworks from cognitive science (Muñoz 2010). On the topic of translation difficulty, two lines of research can be identified in the literature: (1) difficulties in human translation; (2) difficulties in machine translation and post-editing. In a way, they can be situated within the broader framework of cognitive science.

7.4 Human Translation

Two essential questions in translation difficulty research are what makes a text difficult to translate and how to measure and predict the difficulty degree of a translation task. The two questions are complementary, and developing reliable measurement techniques can help advance our understanding of translation difficulty as well as translation processes. Dragsted (2004), for example, 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 is an important variable in translation process research.

Sources of translation difficulty can be divided into two groups: task (i.e., translation) factors and translator factors (Sun 2015). Translation factors include readability (or reading comprehension) problems and translation-specific (or reverbalization) problems, while translator factors concern translation competence (or "ability variables" such as intelligence, aptitude, cognitive style, and working memory capacity), which is more permanent, and affection (or "affective variables" such as confidence, motivation, and anxiety), which is more susceptible to change (Robinson 2001, p. 32). Both groups of factors influence a translator's perception of task difficulty.

In the following three subsections, Sect. 7.4.1 is basically from the perspective of translation-specific problems (or target text characteristics), Sect. 7.4.2 from readability (or source text characteristics), while Sect. 7.4.3 from translator factors.

7.4.1 Choice Network Analysis

Campbell and Hale are pioneers in the empirical exploration of translation difficulty. Campbell and Hale (1999) 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, and explored universal translation difficulties as well as language-specific difficulties. Campbell (1999) found that the source text can

be an independent source of translation difficulty and that a substantial proportion of the items can be equally difficult to translate into typologically different languages.

The way Campbell and Hale assessed the difficulty of a source text was Choice Network Analysis (Campbell 2000), that is, to count the number of different renditions for specific items in that text made by multiple translators. Their rationale was that "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" (Hale and Campbell 2002, p. 15). This rationale has been found problematic. For example, O'Brien (2004) points out that if all translators produce the same solution, the cognitive effort involved might not be less than that required when the translators produce varying target texts; Sun (2015) notes that we cannot assume that the translators are faced with the same number of options in translation, as poor translators usually have fewer (no even none) options than do good translators.

Nonetheless, Dragsted's (2012) empirical study provides evidence for Choice Network Analysis. She found that target text variation was a reliable predictor of difficulty indicators observable in process data, although it was not certain whether high target text variation across participants implied that each individual translator considered several solutions. This finding deserves further exploration. It seems that there are two reasons to explain why Choice Network Analysis may work under some circumstances.

One reason (or source of difficulty in translation) concerns equivalence at a series of levels, especially at the word level and above-word level (see Baker 2011). Non-equivalence, one-to-several equivalence, and one-to-part equivalence situations can create difficulty for translators, especially for novice translators. For instance, the word "presentation" (as in "Students will give a presentation") has no one-to-one equivalent in Mandarin Chinese; translators would have to select from the synonyms like lecture, report, talk, or speech. This would require more cognitive effort on the part of translators, and also create considerable target text variation. It has been found in psychology that the more translations a word has, the lower the semantic similarity of the translation pair (Tokowicz et al. 2002). The translation ambiguity phenomenon is relatively common, especially in some genres such as philosophical writings. In *Dictionary of Untranslatables* (Cassin 2014), over 350 words (e.g., agency, actor) in various languages are explained; one may find that every language expresses a concept with a difference (see also Schleiermacher 2012).

The other reason involves literal translation as a universal initial default strategy in translation, which is related to the Literal Translation Hypothesis. An oft-cited discussion about this hypothesis is as follows:

> The translator begins his search for translation equivalence from formal correspondence, and it is only when the identical-meaning formal correspondent is either not available or not able to ensure equivalence that he resorts to formal correspondents with not-quite-identical meanings or to structural and semantic shifts which destroy formal correspondence altogether. (Ivir 1981, p. 58)

Over the years, translation process researchers seem to have found some experimental evidence in favor of this hypothesis. Englund Dimitrova (2005) observed that translators may use literal translations as provisional solutions in order to minimize cognitive effort, and "there was a tendency for syntactic revisions to result in structures that were more distant from the structure in the ST than the first version chosen" (p. 121). Tirkkonen-Condit (2005) found that the tendency to translate word for word shows in novices as well as professional translators. She argued that literal translation as a default rendering procedure is triggered through automatic processes, and it goes on until interrupted by a monitor that alerts about a problem in the outcome and triggers off conscious decision-making to solve the problem. Balling et al. (2014), on the basis of three eye-tracking experiments, conclude that literal translation may be a universal initial default strategy in translation. Schaeffer and Carl (2014) also found supporting evidence that more translation choices lead to longer reading and processing time.

When translating, if a literal translation is an acceptable solution, translators do not have to exert much cognitive effort and target text variation would obviously be small. If translators have to do syntactic reordering and proceed to less literal ones, the translation would involve more cognitive effort, and also translation competence comes into play. As a result, target text variation would be greater. This means that target text variation and translation difficulty may correlate under certain circumstances. Of course, there are other factors that may affect the process. For example, finding a literal translation may not be equally easy for the translators (Schaeffer and Carl 2013). In addition, Choice Network Analysis entails multiple participants, and is not for measuring translation difficulty for one translator.

7.4.2 A Readability Perspective

Reading comprehension and readability are important topics in reading research. For example, Gray and Leary (1935) in their book *What Makes a Book Readable* presented a comprehensive empirical study, and found that the number of different words in a text, the number of prepositional phrases contained in the text, and the proportionate occurrence of polysyllables bear a significant relationship to the difficulty a reader experiences in reading (pp. 8–9). Since the 1920s, researchers have been working on readability formulas for measuring text readability, and, to date, have published over 200 formulas (e.g., Flesch Reading Ease formula, the Flesch-Kincaid Readability test, the Dale-Chall formula) (Klare 1984). It has been found that vocabulary difficulty and sentence length are the strongest indexes of readability, and other predictor variables add little to the overall predictions of reading difficulty (Chall and Dale 1995). An explanation for this is that many predictor variables correlate with each other. For example, 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*.

However, it should be mentioned that reading difficulty often comes from the ideas rather than the words or sentences. The reason readability formulas work is that difficult passages that express difficult, abstract ideas tend to contain hard words, and vice versa (Rayner and Pollatsek 1989, p. 319). Cheating, that is, "trying to beat the formulas by artificially chopping sentences in half and selecting any short word to replace a long word", may not change the true readability much (Fry 1988, p. 77). This may help explain O'Brien's (2010) empirical finding that the application of controlled language rules increased reading ease only marginally and only for the text identified as "difficult" by the readability formula.

As mentioned earlier, readability (or reading comprehension) is one of the two translation factors that cause translation difficulty. Also, readability-based measurements are objective and consequently can be performed automatically. For these reasons, several translation difficulty researchers have turned to readability formulas for a possible solution.

In an effort to find texts of various translation difficulty levels for experimental purposes, 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), and argues that these objective measures can help us gauge the degree of difficulty of some types of text. Despite this, he warned that the "readability indices cannot give us conclusive evidence of how difficult a translator perceives a text to be" (ibid., p. 67). Mishra et al. (2013), claimed that translation difficulty was mainly caused by three features: sentence length, degree of polysemy of a sentence (i.e., the sum of senses possessed by each word in the WordNet normalized by the sentence length), and sentence structural complexity (i.e., the total length of dependency links in the dependency structure of the sentence). Their experiment, which was based on 80 sentence translations from English into Spanish, Danish and Hindi by at least 2 translators, established a relationship between these three sentential properties and the Translation Difficulty Index, measured as gaze time and fixation count using eye tracking (see Sect. 7.5.2.3 below for meaning). This sentence-based design for measuring translation difficulty might lead to different results than a text-based design.

Most studies on translation difficulty use short texts as test materials. Liu and Chiu (2009) aimed at identifying indicators that may be used to predict source material difficulty for consecutive interpreting, and used four methods to estimate and predict the difficulty of three non-technical source texts, that is, the Flesch Reading Ease formula, information density, new concept density, and expert judgment. They found that these measures all failed statistically in predicting source material difficulty, possibly due to the very small sample size of source texts ($N = 3$). Sun and Shreve (2014) tried to find a method to measure difficulty in a translation task, and found that translation difficulty level and text readability were negatively and weakly related, which means that a text's readability only partially accounted for its translation difficulty level. A post-translation questionnaire survey in their study showed that 77% of over 600 responses pointed to reverbalization in the target language as more difficult than source text comprehension. The survey result was

supported by Hvelplund's (2011) finding that translators allocate more cognitive resources to target text processing than to source text processing as indicated, for instance, by processing time and processing load (i.e., eye-fixation duration). This implies that a readability-based approach may not work for translation difficulty measurement.

7.4.3 Workload Measures and Related Research

Techniques for measuring mental workload can be classified into three major categories: (1) subjective measures, (2) performance measures, and (3) physiological measures (see Sun, 2015). The baseline measure, according to Jex's (1988, p. 14), is the individual's subjective workload evaluation in each task, against which all objective measures must be calibrated.

Performance measures derive an index of workload from some aspect of the participant's behavior or activity, and two commonly used workload indicators are speed (i.e., time-on-task) and accuracy (i.e., number of errors). Alves et al. (2014) in a study used pause duration, drafting time, and number of renditions in microunits as indicators of effort in the translation process. Taking participants' subjective evaluations of translation difficulty with NASA-TLX as the baseline measure, Sun and Shreve (2014) found that translation quality score was an unreliable indicator of translation difficulty level, while time-on-task was significantly, but weakly, related to translation difficulty level. This means that performance measures may not be sensitive to workload manipulations, partly because translation involves solving ill-defined problems (Englund Dimitrova 2005).

As a physiological measure, pupillary responses have been used as an indicator of cognitive load (e.g., Beatty 1982). However, Hvelplund (2011) in his study did not observe a strong relationship between cognitive load and pupillary response in translation. He assumed that larger pupil sizes would reflect heavier cognitive load (p. 71), but found that for students, "pupils were smaller during translation of complex text than during translation of less complex text" (p. 206).

7.5 Post-editing of Machine Translation

Post-editing of machine translation (MT), which "involves a human editor revising an MT output up to an acceptable level of quality" (Kit and Wong 2015), has recently emerged as a major trend in the language industry. According to a survey by DePalma and Hegde (2010) among nearly 1000 language service providers around the world, two-fifths (41.2%) claimed to offer post-edited machine translation. There are a few reasons for the emergence of post-editing. The major one is that translation buyers are increasingly turning to machine translation and post-editing in response to surging content volumes and demand for faster turnaround times

(Common Sense Advisory 2014). The second reason pertains to the change in translation buyers' expectations with regard to the type and quality of translated material (Allen 2003). High-quality translation is expensive and is not always needed. The third reason involves the increasing quality of machine translation output and the wide availability of computer-aided translation (CAT) tools (e.g., SDL Trados, Memsource), which often offer access to third-party machine translation engines (e.g., Google Translate, Microsoft Translator) via an application programming interface (API) and combine computer-aided human translation with post-editing. Translation Automation User Society (TAUS 2014) predicts that post-editing may "overtake translation memory leveraging as the primary production process" in the language industry.

Post-editing is different from human translation in several aspects. In terms of requirements, for example, according to the post-editing guidelines by TAUS (2010), post-editors are expected to "[u]se as much of the raw MT output as possible", and "[e]nsure that no information has been accidentally added or omitted". Vasconcellos (1987) compared post-editing with traditional human revision, and noted that with revision the detection of errors is a discovery process (for e.g., mistranslations, omissions) whereas post-editing is an ongoing exercise of adjusting relatively predictable and recurring difficulties. Thus, post-editing poses specific problems to translators, and prompts strategies different from those used in human translation.

In recent years, there has been an increased interest in the impact of post-editing on cognitive processes (e.g., O'Brien et al. 2014). The factors involved include productivity gains, cognitive effort, impact on quality, quality evaluation and estimation, among others (e.g., Arenas 2014). In post-editing, cognitive effort is largely determined by two main criteria: (1) the quality of the MT raw output; (2) the expected end quality of the translation (TAUS 2010). Generally speaking, the higher the quality of MT output, the less human effort needed for post-editing and hence the higher productivity (Kit and Wong 2015). The expected quality of the final translation can be "good enough" or "similar or equal to human translation", and different quality expectations require different guidelines (TAUS 2010). For example, if the client expects a "good enough" translation, then there is no need to implement corrections or restructure sentences simply to improve the style of the text (ibid.). As these factors are usually discussed under the heading of MT evaluation, we will discuss MT evaluation in the next section in order to put things into perspective.

7.5.1 Machine Translation Evaluation

MT evaluation is intended for assessing the effectiveness and usefulness of existing MT systems and for optimizing their performance, and at the core of evaluation is the quality of MT output (Dorr et al. 2011; Kit and Wong 2015). The notion of quality is context dependent, and its evaluation is influenced by purpose, criteria, text type, and other factors. For this reason, different types of manual (or human) and automatic

evaluation measures have been developed, and with regard to their specific classification there are various proposals.

According to Kit and Wong (2015), manual evaluation of MT output relies on users' subjective judgments and experiences, and entails two aspects: intrinsic and extrinsic. Intrinsic measures focus on judgment of language quality, and include quality assessment, translation ranking, and error analysis; extrinsic measures seek to test the usability of MT output with respect to a specific task, and involve tasks such as information extraction, comprehension test (e.g., cloze test), and post-editing. Automatic evaluation of MT output involves the use of quantitative metrics without human intervention, and includes text similarity metrics and quality estimation.

Automatic measures have been developed to overcome the drawbacks of manual evaluation, such as costliness, subjectivity, inconsistency and slowness, and aim to provide an objective and cost-effective means for MT evaluation. Most of them are text similarity metrics; they judge the quality of MT output by comparing the output against a set of human reference translations of the same source sentences. Among the automatic metrics (see e.g., Dorr et al. 2011), BLEU (Bilingual Evaluation Understudy) (Papineni et al. 2002) is one of the most influential, and its central idea is that "[t]he closer a machine translation is to a professional human translation, the better it is" (p. 311). It counts the *n*-gram (or word sequences) matches between the MT output and human reference translations; the more the matches, the better the MT output is.

Different from text similarity metrics, quality estimation is intended to predict quality of MT output without reference to any human translation. Its rationale is that the "quality of MT output is, to a certain extent, determined by a number of features of the source text and source/target language" (Kit and Wong 2015, p. 230). A study by Felice and Specia (2012) indicated that linguistic features (e.g., number, percentage and ratio of content words and function words, percentage of nouns, verbs and pronouns in the sentence) are complementary to shallow features (e.g., sentence length ratios, type/token ratio variations) for building a quality estimation system. Potential uses of quality estimation include filtering out poor sentence-level translations for human post-editing and selecting the best candidate translation from multiple MT systems (Specia et al. 2010).

Despite being perceived as subjective, manual evaluation has been used as a baseline against which automatic evaluation measures are judged (e.g., Dorr et al. 2011). A strong correlation between automatic evaluation scores and human judgments indicates that the performance of an automatic evaluation metric is satisfactory. BLEU, for example, has a high correlation with human judgments of quality. Quality assessment involves human evaluators who are asked to rate on a five or seven point scale the quality of a translation (normally presented sentence by sentence) in terms of certain characteristics such as fluency and adequacy, and the average score on all the sentences and evaluators is the final score of an MT system (Kit and Wong 2015; Liu and Zhang 2015). Recent studies have shown that judgments of fluency and adequacy are closely related, which may indicate that humans evaluators have difficulty in distinguishing the two criteria (Koehn 2010,

p. 220). Compared with scoring-based quality assessment, translation ranking is often easier. It entails ranking a number of candidate translations of the same sentence from best to worst or picking a preferred version after a pairwise comparison. Since 2008, translation ranking has become the official human evaluation method in the statistical MT workshops organized by the Association for Computational Linguistics (Callison-Burch et al. 2008; Kit and Wong 2015, p. 222). Error analysis is a qualitative process, and it involves identifying MT errors and estimating the amount of work of post-editing, which will be discussed in detail in the next section.

7.5.2 Post-editing Effort Measurement

Post-editing effort can be used to evaluate the quality of machine translation (e.g., Aziz et al. 2013) and develop a suitable pricing model for post-editing. According to a survey by Common Sense Advisory, post-editing pricing ranges widely "from the unedited but optimized, direct-to-publish price of US$0.03 per word to equivalent human rates of US$0.25" (DePalma and Kelly 2009, p. 25). Hence the need for developing metrics for measuring post-editing effort and eventually for predicting post-editing effort.

Krings (2001) in his comprehensive study identified three dimensions of post-editing effort: temporal effort, technical effort, and cognitive effort. Temporal effort refers to time on task (so strictly speaking, it is not an "effort"), while technical effort consists of deletion, insertion, and reordering operations in post-editing. Cognitive effort involves "the type and extent of those cognitive processes that must be activated in order to remedy a given deficiency" (ibid., p. 179) in MT output. According to Krings (2001, pp. 178–182), both temporal effort and technical operations are determined by cognitive effort, which inevitably is the research focus.

In the recent decade, research efforts in measuring cognitive effort in post-editing have been made along three intersecting lines: textual characteristics, characteristics of the translator/post-editor, and workload measures. The first two are causal factors whereas workload measures are effect factors (Meshkati 1988).

7.5.2.1 Characteristics of the Translator/Post-editor

Vieira (2014) investigated the role of individual factors including translators' working memory capacity (WMC) and source language (SL) proficiency in predicting cognitive effort in post-editing, and observed a relationship between WMC and post-editing productivity. This merits further attention. Working memory is usually believed to be a system that combines temporary storage and executive processing in order to help perform complex cognitive activities (Baddeley et al. 2015). As cognitive load is often defined as "the demand for working memory resources

required for achieving goals of specific cognitive activities in certain situations" (Kalyuga 2009, p. 35), working memory is closely related to difficulty.

WMC is an important individual-differences variable for understanding variations in human behavior, and WM span tasks, such as counting span, operation span, and reading span, have been shown to be reliable and valid measures of WMC (Conway et al. 2005). These complex span tasks use "serial recall as a measure of how much [a] person can hold in memory while also performing a secondary task" (Ilkowska and Engle 2010, p. 298). For example, in Daneman and Carpenter's (1980) seminal study, participants read aloud a series of sentences and then recalled the final word of each sentence; the number of final words recalled was the reading span, which varied from two to five. Over the years, it has been found that WMC span measures can predict a broad range of lower-order and higher-order cognitive capabilities, including language comprehension, reasoning, and general fluid intelligence, and that people high in WMC usually outperform those low in WMC in cognitive tasks (see Engle and Kane 2004; Orzechowski 2010). Hummel (2002), for example, noted a significant relation between WMC measured by an L2 reading span task and L2 proficiency. In a translation-related study, Silveira (2011) found that WMC interferes positively in participants' accuracy in the translation task, though not in a significant way in participant's response time.

The strong correlation between WMC scores and higher-order cognition (hence the predictive utility of WMC), however, does not necessarily imply a cause-effect relationship between the two. There have been various explanations and hypotheses regarding the relationship (see Engle and Kane, 2004). One explanation is that WMC scores and higher-order cognition both reflect individual differences in speed of processing. Engle and Kane (2004) argue against it based on their opinion that WMC measures fundamentally tap an attention-control capability. Their argument makes sense in view of the mixed results (e.g., very weak correlations in Conway et al. 2002) reported in literature about the relationship between speed and WMC span measures. Daneman and Carpenter (1980) attributed individual differences in reading span to the chunking process: "the more concepts there are to be organized into a single chunk, the more working memory will be implicated" (p. 464). In other words, differences in the reading span were caused by differences in reading skills. In contrast, Turner and Engle (1989) suggested that WM may be an individual characteristic independent of the nature of the task (e.g., reading, writing), and differences in reading skills were caused by differences in the reading span (see also Engle and Kane 2004).

WM is an active research field. Further discussion of it is out of the scope of this article. Suffice it to say that WM is a central construct in cognitive science, and we believe it is related to other constructs and concepts in translation process research, such as attention, pause, automatization, practice, experience, expertise, and translation competence. It has been noted that the differences between post-editors may be much larger than the difference between machine translation systems (Koehn and Germann 2014, p. 45).

7.5.2.2 Textual Characteristics

Just like the aforementioned readability perspective, textual characteristics of the source text and MT output are believed to be associated with post-editing effort. One of the research goals is to recognize those textual characteristics that can predict PE effort (Vieira 2014). In order to do that, researchers need to identify MT errors and negative translatability indicators (NTIs), the presence of which are supposed to increase PE effort (e.g., O'Brien 2007a). Nonetheless, it should be noted that some source-text features that are normally counted among NTIs may cause increased cognitive processing while those that are usually identified as NTIs (e.g., proper nouns, abbreviations, punctuation problems) may not put demands on cognitive processing (O'Brien 2004, 2005).

There are several classifications of MT errors (see Mesa-Lao 2013 for an overview), which vary according to the type of MT engine, language pair, direction, genre and domain. Wisniewski et al. (2014) found, based on a corpus of English-French automatic translations accompanied with post-edited versions, annotated with PE error labels, that lexical errors accounted for 22%, morphological errors 10%, syntax errors 41%, semantic errors 12%, format errors 5%, and others 10%. Aziz et al. (2014, p. 190) observed that in the English-Spanish PE, production units (PUs, i.e., sequences of successive keystrokes that produce a coherent passage of text) involving verbs tended to be slightly more time-consuming while PUs related to nouns required slightly more typing. Vieira (2014) found that ST prepositional phrases and sentence-level type-token ratio had a significant relationship with cognitive effort in French-English PE, although the effects of source-text linguistic features were small.

In recent years, considerable efforts have been made towards automatic MT error identification and automatic post-editing. For example, through experiments performed on English-to-Brazilian Portuguese MT, Martins and Caseli (2015) found it possible to use the decision tree algorithm to identify wrong segments with around 77% precision and recall. Rosa (2014) developed Depfix, a system for automatic post-editing of phrase-based English-Czech machine translation outputs.

7.5.2.3 Measures of Post-editing Effort

Identifying indices and indicators deemed to reflect or measure how much effort post-editing poses or may pose has been a growing topic of interest. Krings (2001) employed Think-Aloud Protocols (TAP), where post-editors were asked to verbalize their thoughts in a steady stream while performing post-editing in order to investigate their cognitive effort. He made some interesting discoveries, such as verbalization effort (i.e., verbalization volume, as an indicator of cognitive effort) during post-editing of poor MT sentences was about twice as high as that for machine translations evaluated as good (Krings 2001, p. 291). As TAP is a method more qualitative than quantitative (Sun 2011) and some "thoughts pass through the

mind more quickly than they can be verbalized" (Krings 2001, p. 2), calculating the volume and density of verbalizations may not be an ideal method for measuring post-editing effort. O'Brien (2005) proposed the combined use of Campbell's Choice Network Analysis and Translog (to focus on pauses and hesitations) for measuring post-editing effort. Besides these methods, temporal measures, eye tracking, subjective scales, automatic and semi-automatic metrics are often used in the measurement. They can be grouped into the aforementioned three categories: subjective, performance, and physiological measures, and are reviewed in the following paragraphs.

(1) ***Temporal measures***

According to Krings (2001), temporal effort "constitutes the most externally visible indicator of post-editing effort and the most important criterion for determining the economic viability of machine translation" (p. 182), and it is determined by both cognitive and technical efforts. The question is whether temporal effort can be an index of cognitive effort.

Koponen et al. (2012) suggested post-editing time (and its normalized version, seconds per word) as a way to assess post-editing effort, and their experiments indicated that time could be a good metric for understanding post-editing effort. In a study comparing the effort of post-editing English-Portuguese subtitles translated using MT and translation memory systems, Sousa et al. (2011) found a good correlation between their objective way of measuring post-editing effort (i.e., time) and subjective evaluation scores.

However, studies by De Almeida (2013) and Arenas (2014) reported no correlation between the participants' levels of experience and the total time taken to complete the post-editing task, and revealed a complex relationship between PE effort, PE performance, and previous experience. These studies did not address directly the correlation between PE time and PE effort. Nonetheless, they cast into doubt whether PE time alone is a robust measure of PE effort (Vieira 2014) and call for further studies.

(2) ***Pause analysis***

Pauses have been used as indicators of cognitive processing in research on translation (e.g., Lacruz and Shreve 2014) as well as on speech production and writing (e.g., Schilperoord 1996). They help reduce the load on working memory, and would not occur provided that "(1) sufficient cognitive resources are dedicated to the [. . .] process; (2) the [process] is sufficiently automated (concerning syntactic, lexical and graphomotor processes); and (3) domain knowledge is sufficiently activated for it to be retrieved at a lesser cost" (Alamargot et al. 2007, p. 16). Pause analysis often focuses on three dimensions: pause duration, position, and proportion. In writing studies, for example, pauses are usually interpreted on the basis of four assumptions, one of which is that "pause duration varies as a function of the complexity of the processes engaged in" (Foulin, from Alamargot et al. 2007, p. 14).

In the field of post-editing, O'Brien (2006) investigated the relationship between pauses (recorded using Translog) and PE cognitive effort, which was indicated by differences in negative translatability indicators in the source text and Choice Network Analysis. She found little correspondence between pause duration and editing of the "difficult" element, though those "difficult" elements identified by Choice Network Analysis were always preceded by a pause.

Based on observations in a case study, Lacruz et al. (2012) introduced the average pause ratio metric, which was found to be sensitive to the number and duration of pauses, and noted that APR could be a potentially valid measure of cognitive demand. In their follow-up study involving three participants, Lacruz and Shreve (2014) found that the behavioral metrics of average pause ratio and pause to word ratio appeared to be strongly associated with PE cognitive effort, which was measured indirectly by computing the number of complete editing events (i.e., event to word ratio, or EWR) in the TT segment from the keystroke log report. A complete editing event refers to "a sequence of actions leading to linguistically coherent and complete output" (ibid., p. 250), and the assumption of its use was that "each editing event resulted from a coherent expenditure of cognitive effort in post-editing the MT segment" (ibid., pp. 251–252) and more events would indicate higher cognitive effort. Another assumption was that all complete editing events would require the same amount of cognitive effort, which, as mentioned in their article, was problematic. Also, when computing EWR, the judgment of what constitutes a complete editing event is "to some extent subjective" (ibid., p. 269), and this makes the time-consuming manual analysis of keystroke logs difficult to automate.

As noted in both O'Brien's (2006) and Lacruz and Shreve's (2014) studies, the patterns of pause activity vary from one individual to another. Koponen et al. (2012) observed that post-editors adopted different editing trategies. For example,

> [S]ome editors maximize the use of MT words and cut-paste operations for reordering, while others appear to prefer writing out the whole corrected passage and then deleting MT words even when they are the same. . .[S]ome editors spend their time planning the corrections first and proceeding in order while others revise their own corrections and move around in the sentence. (Koponen et al. 2012, p. 19)

These different strategies would certainly impact the pausing behavior. Further studies in this direction preferably need to adopt a within-subject or longitudinal design. Nevertheless, we agree with O'Brien (2006) that pauses on their own probably are not a robust measure of PE effort. As an online method, pause analysis would better be used to identify problems in translation or PE and determine the workload levels associated with those problems.

(3) ***Eye tracking***

The method of using an eye tracker to record eye movements and pupil size variation has been employed to investigate various cognitive processes in reading (e.g., Rayner 1998), writing (e.g., Alamargot et al. 2006), usability testing (e.g., Poole and Ball 2006), translation (e.g., Gopferich et al. 2008; O'Brien 2007b), and other fields. Common eye-tracking metrics include gaze time, fixation counts,

fixation durations, pupil dilation, blink rate, and scanpath similarity (for their meaning, see O'Brien 2011, pp. 238–241).

A fundamental assumption in eye-tracking research is the eye-mind hypothesis (Just and Carpenter 1976, 1980), which posits that "the locus of the eye fixations reflects what is being internally processed" (1976, p. 471) and "the eye remains fixated on a word as long as the word is being processed" (1980, p. 330). Thus gaze time directly indicates the time it takes to process a fixated word. Of course, this hypothesis may not be valid under certain circumstances, e.g., during mindless reading, in which the eyes continue moving across the page (or screen) even though the mind is thinking about something unrelated to the text (Reichle et al. 2010). For this reason, Just and Carpenter (1976) specified several conditions for their hypothesis to be valid, e.g., asking the participant to work accurately but quickly and specifying the exact task goals (also see Goldberg and Wichansky 2003).

Eye-tracking metrics have been used to measure cognitive load (cf. Tatler et al. 2014), based on such assumptions as: (1) longer gaze time (i.e., the sum of all fixation durations) corresponds to an increased level of cognitive processing; (2) fixation count (i.e., the number of fixations) is related to the number of components that an individual is required to process (Fiedler et al. 2012, p. 26). These assumptions were supported by Doherty et al.'s (2010) study testing the validity of eye tracking as a means of evaluating MT output, in which they found that gaze time and fixation count correlated reasonably well with human evaluation of MT output. However, they observed that average fixation duration and pupil dilations were not reliable indicators of reading difficulty for MT output, which corroborated the finding of Sharmin et al. (2008) that fixations in translation tasks were more frequent if the source text was complex, but not longer.

In translation process research, eye tracking is usually used together with keystroke logging (e.g., Translog, Inputlog), and is supposed to provide data "all through a translation task without interruption or with few interruptions and to be able to 'fill' most of the pauses in keystroke activity with interesting data" (Jakobsen 2011, p. 41). A notable effort in this regard is the EU-funded Eye-to-IT project.

(4) ***Evaluation metrics***

As mentioned in Sect. 7.5.1, there are manual and automatic MT evaluation measures. Automatic measures usually entail reference translations, and one common metric is edit distance, which refers to the minimum number of modifications (i.e., deletions, insertions, substitutions) required to transform an MT output into a reference translation. Translation Edit Rate (TER) (Snover et al. 2006) is a major edit-distance-based metric, and it correlates reasonably well with human judgments of MT quality. Reference translations can be done by professional translators (used in metrics like BLEU and TER) or created by post-editing the MT output. In the latter case, a popular metric is human-targeted Translation Edit Rate (HTER) (Snover et al. 2006), which guarantees only minimum number of edits necessary to transform an MT output into a fluent and adequate translation; it was found (ibid.) that HTER correlates with human judgments better than BLEU, TER, or METEOR (Banerjee and Lavie 2005), another major automatic evaluation metric.

Vieira (2014) in his study found that METEOR (Denkowski and Lavie 2011) was significantly correlated with all measures of PE cognitive effort considered, which included ST and MT-output characteristics and individual factors, especially for longer sentences. However, such metrics are measures of technical effort and do not directly measure cognitive effort. Koponen (2012) investigated the relationship between cognitive and technical aspects of post-editing effort by comparing translators' perceived PE effort (as indicated by scores on a 1–5 scale) to actual edits made (measured by HTER) and found they did not always correlate with each other. Some edits are more difficult than others; certain types of errors require great cognitive effort although they involve few edits, and vice versa (Koponen et al. 2012).

(5) ***Subjective scales***

Human perceptions of PE effort have been used often in studies measuring PE cognitive effort. In Specia's study (2011), for instance, after post-editing each sentence, translators were asked to score the original translation according to its post-editing effort on a 4-point scale (with 1 being requiring complete retranslation and 4 being fitting for purpose). For the purpose of measuring cognitive effort, Vieira (2014) used Paas's (1992) 9-point Likert scale together with average fixation duration and fixation count.

Towards the use of subjective scales, De Waard and Lewis-Evans (2014) expressed reservations and argued that since subjective ratings have no actual absolute reference, ratings between conditions can only be compared in within-subject designs, and that the variation of workload during task performance cannot be reflected in one rating. About the first point, the assumption in Sect. 7.6.2 may provide an explanation for why between-subject designs can also be used although within-subject designs are probably better in most cases. The second point makes sense and that is why many researchers choose to use subjective scales together with eye tracking and/or pause analysis.

This section has reviewed the uses of temporal measures, pause analysis, eye tracking, evaluation metrics, and subjective scales in PE effort measurement. Their suitability for this purpose aside, adoption of these methods involves a trade-off between granularity of analysis and volume of analysis (Moran and Lewis 2011). For example, analysis of post-editing using eye tracking usually involves fewer test sentences compared with other methods, but can draw on more highly granular data.

In PE studies, PE platforms are often adopted to facilitate research in e.g., keystroke analysis, temporal measurement, edit distance calculation, error annotation, or subjective rating. Such platforms include Blast (Stymne 2011), CASMACAT (Ortiz-Martínez et al. 2012), PET (Aziz and Specia 2012), TransCenter (Denkowski and Lavie 2012), and others.

7.6 Assumptions in Translation Difficulty Research

As mentioned in the previous section, assumptions are inherent in pause analysis and eye tracking research. According to Kanazawa (1998, p. 196), a scientific theory consists of two major parts: assumptions and hypotheses, and assumptions are "universal axiomatic statements about some part of the empirical world". Assumptions can be implicit or explicit, and it is important that they are made explicit, especially in research methods used to test theories. Nkwake (2013, pp. 107–109) proposes five categories of assumptions on a continuum of explication: (1) very ambiguously tacit assumptions, (2) tacit but more obvious assumptions, (3) informally, explicit assumptions, (4) assumptions that are made explicit, and (5) explicit and tested assumptions. In this section, we make explicit some assumptions in translation difficulty research.

7.6.1 Assumption of Linearity

The assumption of linearity is that there is a straight-line relationship between the two variables of interest (e.g., Onwuegbuzie and Daniel 1999). It is an important assumption in parametric statistics involving two or more continuous variables (e.g., ANOVA, linear regression), and researchers generally assume linear relationships in their data (Nimon 2012). However, linearity is not guaranteed and should be validated (ibid, p. 4).

In the workload-related literature, researchers (e.g., Cassenti et al. 2013; O'Donnell and Eggemeier 1986) have proposed an overall nonlinear relationship between workload and performance. Fig. 7.1 is the hypothesized workload-performance curve by Cassenti and Kelley (2006; see Cassenti et al. 2013).

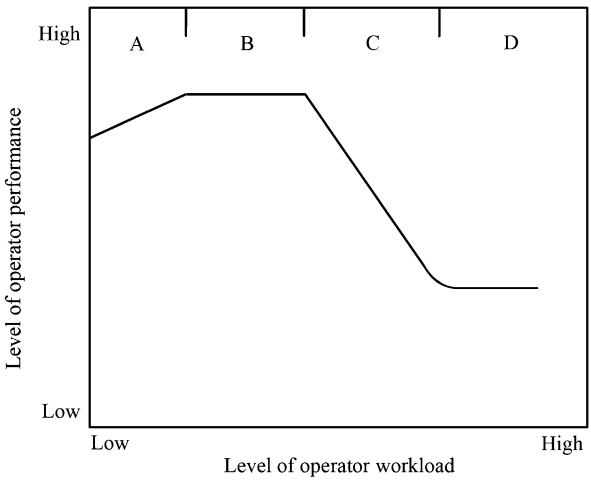

Fig. 7.1 Cassenti and Kelley's hypothesized workload-performance curve

This curve has four linear segments. For easy tasks, increased workload may lead to improved performance, or is not accompanied by variations in 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 and Eggemeier 1986). Correspondingly, there may be some dissociation of performance and subjective measures under certain conditions, especially when the workload is very low (floor effect) or very high (ceiling effect) (Vidulich and Tsang 2012, p. 259). The implications for translation difficulty research are that the experiment tasks should be moderately difficult for the participants.

7.6.2 The Same Ranking Assumption

This assumption has two meanings: (1) If a novice believes Passage A is more difficult to translate than Passage B, it is so for a professional translator (unless she works in the domain of Passage A for a long time); (2) If Passage A is more difficult to translate than Passage B for a translator, it will remain so for her (unless she works in the domain of Passage A for a long time). This assumption is based on the finding that "translation does not become easier with growing experience and expertise" (Sirén and Hakkarainen 2002, p. 71). A ranking is valid as long as "there is a single criterion [by] which the objects are evaluated, or the objects map to a linear scale" (Busse and Buhmann 2011, p. 220). This implies that the assumption of linearity is a prerequisite for this assumption.

In a relevant study, Tomporowski (2003) found that training improved participants' performance on each of the three cognitive tasks in his experiment; however, training on one task (i.e., Paced Auditory Serial Addition Task) did not lead to changes in total NASA-TLX workload ratings, whereas training on the other two tasks (i.e., an attentional-switching task and a response-inhibition task) led to decreased ratings of overall workload. That is, the impact of training on workload ratings may be task dependent. This finding can be explained in terms of the effect of practice on the development of automaticity and expertise. In writing process research, it has been noted that lexicon access and graphomotor execution can be automated with practice, while the processes involved in content generation, such as planning and reviewing, are difficult to become automatic (Alamargot et al. 2007, p. 15). Thus, features of automaticity in the translation process may need separate investigations (Moors and De Houwer 2006).

Obviously, both the linearity assumption and the same ranking assumption require testing. Nonetheless, assumptions are necessarily simplifications and can help "explain phenomena in that part of the empirical world" (Kanazawa 1998, p. 198).

7.7 Discussion and Conclusion

This article has reviewed methods for measuring difficulty in translation and post-editing and relevant research. One major reason for measuring difficulty is to avoid cognitive overload and underload, and help maintain optimal performance. From a pedagogical perspective, one needs deliberate practice in order to become an expert in an area (e.g., translation, post-editing), and one of the conditions is that the task is of appropriate difficulty for the individual (Shreve 2002).

In terms of research methods, it is important to find a valid and reliable method for after-the-fact measurement of difficulty for the individual. It seems that the individual's subjective workload evaluation in each task can serve as a baseline measure in translation and post-editing. Yet, paradoxically, the ultimate objective is to find an objective and automatic way to predict the workload independent of the individuals. Thus, researchers turn to language factors (i.e., source text features, translated/post-edited text features, and their correspondence). A typical research procedure is that researchers count the occurrences of the language factors (or translation errors) in the criterion passages and correlate them with the difficulty values of the criterion passages in order to select the factors that can work as potential indexes of difficulty in translation or post-editing (Anagnostou and Weir 2007, p. 6).

Translation and post-editing are closely related activities. Post-editors are usually translators. In difficulty research, they share some objectives and can draw on each other's research methods and findings. Of course, they have differences in, among others, operations and behavior, translation error patterns, translation quality expectations, and research designs. For example, in a translation task, the quality is usually expected to be as good as possible, whereas in post-editing, post-editors are often instructed to use as much of the raw MT output as possible. Differences in quality expectations naturally lead to differences in perceived workload. Research indicates that the post-editing effort for all segments was lower than the translation effort for those segments (O'Brien 2007a). In terms of test materials, a group of sentences are usually adopted in post-editing studies while texts in translation difficulty research. Reading researchers have found that readers tend to "pause at the location where the difficulty arises. . .regress to an earlier part of the text, or they postpone solving the problem and move on to the next part of the text hoping to find a solution there" (Vonk and Cozijn 2003). In post-editing studies, however, participants are often presented with one sentence at a time, with revisits not being allowed (Vieira 2014, p. 212).

In this review, we have assumed that all the research findings are equally trustworthy. This is not the case, although no one would deny that they are all valuable. Many, if not most, studies have a small sample size with a few participants. Findings from these exploratory studies are difficult to generalize, especially because there are many variables involved, such as text type, domain, language directionality, and professional experience. Hence the need for replication studies. A significant effort in this direction has been the translation process research database (TPR-DB) developed by the Center for Research and Innovation in Translation and Translation

Technology (CRITT) at Copenhagen Business School, which stores Translog-II data from reading, writing, translation, copying and post-editing experiments, and CASMACAT translation sessions from various language combinations (see Carl et al. 2015). There is some way to go before researchers can find an objective way to measure and predict difficulty in translation and post-editing.

Acknowledgment This work was supported by the Young Faculty Research Fund of Beijing Foreign Studies University (Grant No. 2016JT004) and by the Fundamental Research Funds for the Central Universities (Grant No. 2015JJ003).

References

Alamargot, D., Chesnet, D., Dansac, C., & Ros, C. (2006). Eye and pen: A new device for studying reading during writing. *Behavior Research Methods, 38*(2), 287–299.

Alamargot, D., Dansac, C., Chesnet, D., & Fayol, M. (2007). Parallel processing before and after pauses: A combined analysis of graphomotor and eye movements during procedural text production. In M. Torrance, L. Van Waes, & D. Galbraith (Eds.), *Writing and cognition: Research and applications* (pp. 13–29). Amsterdam: Elsevier.

Allen, J. (2003). Post-editing. In H. Somers (Ed.), *Computers and translation: A translator's guide* (pp. 297–318). Amsterdam: John Benjamins.

Alves, F. (2015). Translation process research at the interface. In A. Ferreira & J. W. Schwieter (Eds.), *Psycholinguistic and cognitive inquiries into translation and interpreting* (pp. 17–39). Amsterdam: John Benjamins.

Alves, F., Pagano, A., & da Silva, I. (2014). Effortful text production in translation: A study of grammatical (de)metaphorization drawing on product and process data. *Translation and Interpreting Studies, 9*(1), 25–51.

Anagnostou, N. K., & Weir, G. R. S. (2007). From corpus-based collocation frequencies to readability measure. In G. R. S. Weir & T. Ozasa (Eds.), *Texts, textbooks and readability* (pp. 34–48). Glasgow: University of Stratchclyde Publishing.

Arenas, A. G. (2014). The role of professional experience in post-editing from a quality and productivity perspective. In S. O'Brien, L. W. Balling, M. Carl, M. Simard, & L. Specia (Eds.), *Post-editing of machine translation: Processes and applications* (pp. 51–76). Newcastle: Cambridge Scholars Publishing.

Aziz, W., & Specia, L. (2012). *PET: A standalone tool for assessing machine translation through post-editing*. Paper presented at the Translating and The Computer 34, London.

Aziz, W., Mitkov, R., & Specia, L. (2013). Ranking machine translation systems via post-editing. In I. Habernal & V. Matoušek (Eds.), *Text, speech, and dialogue* (pp. 410–418). London: Springer.

Aziz, W., Koponen, M., & Specia, L. (2014). Sub-sentence level analysis of machine translation post-editing effort. In S. O'Brien, L. W. Balling, M. Carl, M. Simard, & L. Specia (Eds.), *Post-editing of machine translation: Processes and applications* (pp. 170–199). Newcastle: Cambridge Scholars Publishing.

Baddeley, A. D., Eysenck, M. W., & Anderson, M. C. (2015). *Memory* (2nd ed.). London: Psychology Press.

Baker, M. (2011). *In other words: A coursebook on translation* (2nd ed.). New York: Routledge.

Balling, L. W., Hvelplund, K. T., & Sjørup, A. C. (2014). Evidence of parallel processing during translation. *Meta, 59*(2), 234–259.

Banerjee, S., & Lavie, A. (2005). *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments.* Paper presented at the Workshop on Intrinsic and Extrinsic

Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan.
Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276–292.
Bermúdez, J. L. (2014). *Cognitive science: An introduction to the science of the mind* (2nd ed.). Cambridge: Cambridge University Press.
Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica, 134*(3), 330–343.
Bradshaw, J. L. (1968). Load and pupillary changes in continuous processing tasks. *British Journal of Psychology, 59*(3), 265–271.
Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon Press.
Brünken, R. E., Plass, J. L., & Moreno, R. E. (2010). *Current issues and open questions in cognitive load research*. Cambridge: Cambridge University Press.
Busse, L. M., & Buhmann, J. M. (2011). Model-based clustering of inhomogeneous paired comparison data. In M. Pelillo & E. R. Hancock (Eds.), *Similarity-based pattern recognition* (pp. 207–221). Berlin: Springer.
Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2008). Further meta-evaluation of machine translation. In *Proceedings of the third workshop on statistical machine translation* (pp. 70–106). Columbus: Association for Computational Linguistics.
Campbell, S. (1999). A cognitive approach to source text difficulty in translation. *Target, 11*(1), 33–63.
Campbell, S. (2000). Choice network analysis in translation research. In M. Olohan (Ed.), *Intercultural faultlines: Research models in translation studies: Textual and cognitive aspects* (pp. 29–42). Manchester: St. Jerome.
Campbell, S., & Hale, S. (1999). What makes a text difficult to translate? *Refereed Proceedings of the 23rd Annual ALAA Congress*. Retrieved March 1, 2015, from http://www.atinternational.org/forums/archive/index.php/t-887.html
Cara, F. (1999). Cognitive ergonomics. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 130–132). Cambridge: MIT Press.
Carl, M., Bangalore, S., & Schaeffer, M. (2015). *New directions in empirical translation process research: Exploring the CRITT TPR-DB*. New York: Springer.
Cassenti, D. N., & Kelley, T. D. (2006). *Towards the shape of mental workload.* Paper presented at the Human Factors and Ergonomics Society Annual Meeting, Boston, MA.
Cassenti, D. N., Kelley, T. D., & Carlson, R. A. (2013). *Differences in performance with changing mental workload as the basis for an IMPRINT plug-in proposal.* Paper presented at the 22nd Annual Conference on Behavior Representation in Modeling and Simulation, Ottawa, Canada.
Cassin, B. (2014). *Dictionary of untranslatables: A philosophical lexicon*. Princeton: Princeton University Press.
Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Cambridge: Brookline Books.
Charlton, S. G. (2002). Measurement of cognitive states in test and evaluation. In S. G. Charlton & T. G. O'Brien (Eds.), *Handbook of human factors testing and evaluation* (2nd ed., pp. 97–126). Mahwah: Lawrence Erlbaum.
Common Sense Advisory. (2014). Ten concepts and data points to remember in 2014. *MultiLingual, 1*, 37-38.
Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163–183.
Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5), 769–786.
Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450–466.

De Almeida, G. (2013). *Translating the post-editor: An investigation of post-editing changes and correlations with professional experience across two Romance languages.* PhD thesis. Dublin City University, Dublin.

De Waard, D., & Lewis-Evans, B. (2014). Self-report scales alone cannot capture mental workload. *Cognition, Technology & Work, 16*(3), 303–305.

DeKeyser, R. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 313–348). Oxford: Blackwell.

Denkowski, M., & Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the 6th workshop on statistical machine translation* (pp. 85–91). Edinburgh: Association for Computational Linguistics.

Denkowski, M., & Lavie, A. (2012). *TransCenter: Web-based translation research suite*. Retrieved April 1, 2015, from https://www.cs.cmu.edu/~mdenkows/pdf/transcenter-amta2012.pdf

DePalma, D. A., & Hegde, V. (2010). *The market for MT post-editing*. Lowell: Common Sense Advisory.

DePalma, D. A., & Kelly, N. (2009). *The business case for machine translation*. Lowell: Common Sense Advisory.

Doherty, S., O'Brien, S., & Carl, M. (2010). Eye tracking as an MT evaluation technique. *Machine Translation, 24*(1), 1–13.

Dorr, B., Olive, J., McCary, J., & Christianson, C. (2011). Machine translation evaluation and optimization. In J. Olive, C. Christianson, & J. McCary (Eds.), *Handbook of natural language processing and machine translation* (pp. 745–843). New York: Springer.

Dragsted, B. (2004). *Segmentation in translation and translation memory systems: An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process.* PhD thesis. Copenhagen Business School, Frederiksberg, Denmark.

Dragsted, B. (2012). Indicators of difficulty in translation: Correlating product and process data. *Across Languages and Cultures, 13*(1), 81–98.

Embrey, D., Blackett, C., Marsden, P., & Peachey, J. (2006). *Development of a human cognitive workload assessment tool: MCA final report*. Dalton: Human Reliability Associates.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The psychology of learning and motivation* (pp. 145–199). New York: Elsevier.

Englund Dimitrova, B. (2005). *Expertise and explicitation in the translation process*. Amsterdam: John Benjamins.

Felice, M., & Specia, L. (2012). Linguistic features for quality estimation. In *Proceedings of the 7th workshop on statistical machine translation* (pp. 96–103). Montréal: Association for Computational Linguistics.

Fiedler, S., Glöckner, A., & Nicklisch, A. (2012). The influence of social value orientation on information processing in repeated voluntary contribution mechanism games: An eye-tracking analysis. In A. Innocenti & A. Sirigu (Eds.), *Neuroscience and the Economics of Decision Making* (pp. 21–53). London: Routledge.

Frankish, K., & Ramsey, W. (Eds.). (2012). *The Cambridge handbook of cognitive science*. Cambridge: Cambridge University Press.

Freeman, G. L., & Giese, W. J. (1940). The relationship between task difficulty and palmar skin resistance. *The Journal of General Psychology, 23*(1), 217–220.

Freixa, J. (2006). Causes of denominative variation in terminology: A typology proposal. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 12*(1), 51–77.

Fry, E. B. (1988). Writeability: The principles of writing for increased comprehension. In B. L. Zakaluk & S. J. Samuels (Eds.), *Readability: Its past, present, and future* (pp. 77–95). Newark: International Reading Association.

Gallupe, R. B., DeSanctis, G., & Dickson, G. W. (1988). Computer-based support for group problem-finding: An experimental investigation. *MIS Quarterly, 12*(2), 277–296.

Gile, D. (1999). Testing the Effort Models' tightrope hypothesis in simultaneous interpreting – A contribution. *Hermes, 23*, 153–172.
Gile, D. (2009). *Basic concepts and models for interpreter and translator training* (Rev. Ed.). Amsterdam: John Benjamin.
Goldberg, J. H., & Wichansky, A. M. (2003). Eye tracking in usability evaluation: A practitioner's guide. In R. Radach, J. Hyona, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 493–516). Amsterdam: Elsevier.
Göpferich, S., Jakobsen, A. L., & Mees, I. M. (Eds.). (2008). *Looking at eyes: Eye-tracking studies of reading and translation processing*. Copenhagen: Sammfundslitteratur.
Gopher, D. (1994). Analysis and measurement of mental load. In G. d'Ydewalle, P. Eelen, & P. Bertelson (Eds.), *International perspectives on psychological science, Vol. II: The state of the art* (pp. 265–292). East Sussex: Lawrence Erlbaum.
Gray, W. S., & Leary, B. E. (1935). *What makes a book readable*. Chicago: The University of Chicago Press.
Hale, S., & Campbell, S. (2002). The interaction between text difficulty and translation accuracy. *Babel, 48*(1), 14–33.
Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam: North-Holland.
Hummel, K. M. (2002). Second language acquisition and working memory. In F. Fabbro (Ed.), *Advances in the neurolinguistics of bilingualism* (pp. 95–117). Udine: Forum.
Hvelplund, K. T. (2011). *Allocation of cognitive resources in translation: An eye-tracking and key-logging study.* PhD thesis. Copenhagen Business School, Frederiksberg, Denmark.
Ilkowska, M., & Engle, R. W. (2010). Trait and state differences in working memory capacity. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of individual differences in cognition* (pp. 295–320). New York: Springer.
International Ergonomics Association. (2015). *Definition and domains of ergonomics*. Retrieved March 1, 2015, from http://www.iea.cc/whats/
Ivir, V. (1981). Formal correspondence vs. translation equivalence revisited. *Poetics Today, 2*(4), 51–59.
Jakobsen, A. L. (2011). Tracking translators' keystrokes and eye movements with Translog. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research* (pp. 37–55). Amsterdam: John Benjamins.
Jensen, K. T. (2009). Indicators of text complexity. In S. Göpferich, A. L. Jakobsen, & I. M. Mees (Eds.), *Behind the mind: Methods, models and results in translation process research* (pp. 61–80). Amsterdam: John Benjamins.
Jex, H. R. (1988). Measuring mental workload: Problems, progress, and promises. In P. A. Hancock & N. Meshkati (Eds.), *Hman mental workload* (pp. 5–38). Amsterdam: North-Holland.
Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology, 8*(4), 441–480.
Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354.
Kalsbeek, J. W. H., & Sykes, R. N. (1967). Objective measurement of mental load. *Acta Psychologica, 27*, 253–261.
Kalyuga, S. (2009). *Managing cognitive load in adaptive multimedia learning*. Hershey: Information Science Reference.
Kanazawa, S. (1998). In defense of unrealistic assumptions. *Sociological Theory, 16*(2), 193–204.
Karwowski, W. (2012). The discipline of human factors and ergonomics. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 1–37). Hoboken: Wiley.
Kit, C. Y., & Wong, B. T. M. (2015). Evaluation in machine translation and computer-aided translation. In S. W. Chan (Ed.), *Routledge encyclopedia of translation technology* (pp. 213–236). London: Routledge.

Klare, G. R. (1984). Readability. In P. D. Pearson & R. Barr (Eds.), *Handbook of reading research* (pp. 681–744). New York: Longman.

Koehn, P. (2010). *Statistical machine translation*. New York: Cambridge University Press.

Koehn, P., & Germann, U. (2014). *The impact of machine translation quality on human post-editing*. Paper presented at the Workshop on Humans and Computer-Assisted Translation (HaCaT), Gothenburg, Sweden.

Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the 7th Workshop on Statistical Machine Translation* (pp. 181–190). Montreal: Association for Computational Linguistics.

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). *Post-editing time as a measure of cognitive effort.* Paper presented at the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012), San Diego.

Krings, H. P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes*. (G. Koby, G. Shreve, K. Mischerikow & S. Litzer, Trans.). Kent, Ohio: Kent State University Press.

Kuiken, F., & Vedder, I. (2007). Task complexity needs to be distinguished from task difficulty. In M. D. P. GarcíaMayo (Ed.), *Investigating tasks in formal language learning* (pp. 117–135). Clevedon: Multilingual Matters.

Lacruz, I., & Shreve, G. M. (2014). Pauses and cognitive effort in post-editing. In S. O'Brien, L. W. Balling, M. Carl, M. Simard, & L. Specia (Eds.), *Post-editing of machine translation: Processes and applications* (pp. 246–272). Newcastle: Cambridge Scholars Publishing.

Lacruz, I., Shreve, G. M., & Angelone, E. (2012). *Average pause ratio as an indicator of cognitive effort in post-editing: A case study.* Paper presented at the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012), San Diego.

Liu, M., & Chiu, Y.-H. (2009). Assessing source material difficulty for consecutive interpreting: Quantifiable measures and holistic judgment. *Interpreting, 11*(2), 244–266.

Liu, Q., & Zhang, X. (2015). Machine translation: General. In S. W. Chan (Ed.), *Routledge encyclopedia of translation technology* (pp. 105–119). London: Routledge.

Lusk, M. M., & Atkinson, R. K. (2007). Animated pedagogical agents: Does their degree of embodiment impact learning from static or animated worked examples? *Applied Cognitive Psychology, 21*(6), 747–764.

Martins, D. B., & Caseli, H. (2015). Automatic machine translation error identification. *Machine Translation, 29*(1), 1–24.

Mesa-Lao, B. (2013). *Introduction to post-editing–The CasMaCat GUI*. Retrieved March 1, 2015 from http://bridge.cbs.dk/projects/seecat/material/hand-out_post-editing_bmesa-lao.pdf

Meshkati, N. (1988). Toward development of a cohesive model of workload. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 305–314). Amsterdam: North-Holland.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Mishra, A., Bhattacharyya, P., & Carl, M. (2013, August 4–9). *Automatically predicting sentence translation difficulty.* Paper presented at the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria.

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin, 132*(2), 297–326.

Moran, J., & Lewis, D. (2011). *Unobtrusive methods for low-cost manual evaluation of machine translation.* Retrieved April 1, 2015 from http://lodel.irevues.inist.fr/tralogy/index.php?id=141&format=print

Moray, N. (1977). Models and measures of mental workload. In N. Moray (Ed.), *Mental workload: Its theory and measurement* (pp. 13–21). New York: Springer.

Muñoz Martín, R. (2010). Leave no stone unturned: On the development of cognitive translatology. *Translation and Interpreting Studies, 5*(2), 145–162.

Muñoz Martín, R. (2012). Just a matter of scope. *Translation Spaces, 1*(1), 169–188.

Muñoz Martín, R. (2014). A blurred snapshot of advances in translation process research. *MonTI. Special Issue (Minding Translation), 1*, 49–84.

Newell, W. H. (2001). A theory of interdisciplinary studies. *Issues in Integrative Studies, 19*, 1–25.

Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology, 3*, 1–5.

Nkwake, A. M. (2013). *Working with assumptions in international development program evaluation*. New York: Springer.

Nord, C. (2005). *Text analysis in translation: Theory, methodology, and didactic application of a model for translation-oriented text analysis* (2nd ed.). Amsterdam: Rodopi.

O'Brien, S. (2004). *Machine Translatability and Post-Editing Effort: How do they relate?* Paper presented at the 26th Translating and the Computer Conference (ASLIB), London.

O'Brien, S. (2005). Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation, 19*(1), 37–58.

O'Brien, S. (2006). Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Languages and Cultures, 7*(1), 1–21.

O'Brien, S. (2007a). An empirical investigation of temporal and technical post-editing effort. *Translation and Interpreting Studies, 2*(1), 83–136.

O'Brien, S. (2007b). Eye-tracking and translation memory matches. *Perspectives, 14*(3), 185–205.

O'Brien, S. (2010). Controlled language and readability. In G. M. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 143–165). Amsterdam: John Benjamins.

O'Brien, S. (2011). *Cognitive explorations of translation*. London: Continuum.

O'Brien, S., Balling, L. W., Carl, M., Simard, M., & Specia, L. (Eds.). (2014). *Post-editing of machine translation: Processes and applications*. Newcastle: Cambridge Scholars Publishing.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance, Vol. II: Cognitive processes and performance* (pp. 42/41–42–49). New York: Wiley.

Onwuegbuzie, A. J., & Daniel, L. G. (1999, November 17–19). *Uses and misuses of the correlation coefficient.* Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Point Clear, AL.

Ortiz-Martínez, D., Sanchis-Trilles, G., Casacuberta, F., Alabau, V., Vidal, E., Benedı, J.-M … González, J. (2012). *The CASMACAT project: The next generation translator's workbench.* Paper presented at the 7th Jornadas en Tecnologıa del Habla and the 3rd Iberian SLTech Workshop (IberSPEECH), Madrid.

Orzechowski, J. (2010). Working memory capacity and individual differences in higher-level cognition. In G. Matthews & B. Szymura (Eds.), *Handbook of individual differences in cognition* (pp. 353–368). New York: Springer.

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429–434.

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1994a). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*(4), 351–371.

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1994b). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*(1), 122–133.

Paas, F. G. W. C., Ayres, P., & Pachman, M. (2008). Assessment of cognitive load in multimedia learning. In D. H. Robinson & G. Schraw (Eds.), *Assessment of cognitive load in multimedia learning: Theory, methods and applications* (pp. 11–35). Charlotte, NC: Information Age Publishing.

Palumbo, G. (2009). *Key terms in translation studies*. London: Continuum.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318). University of Pennsylvania, Philadelphia: Association for Computational Linguistics.

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research. In C. Ghaoui (Ed.), *Encyclopedia of human computer interaction* (pp. 211–219). London: Idea Group.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.

Rayner, K., & Pollatsek, A. (1989). *Psychology of reading*. Hillsdale: Lawrence Erlbaum.

Redfield, C. L. (1922). Mental levels. *Journal of Education, 95*(8), 214–216.

Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010). Eye movements during mindless reading. *Psychological Science, 21*(9), 1300–1310.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287–318). Cambridge: Cambridge University Press.

Rosa, R. (2014). Depfix, a tool for automatic rule-based post-editing of SMT. *The Prague Bulletin of Mathematical Linguistics, 102*(1), 47–56.

Rost, M. (2006). Areas of research that influence L2 listening instruction. In E. Usó Juan & A. Martínez Flor (Eds.), *Current trends in the development and teaching of the four language skills* (pp. 47–74). Berlin: Mouton de Gruyter.

Schaeffer, M., & Carl, M. (2013). Shared representations and the translation process: A recursive model. *Translation and Interpreting Studies, 8*(2), 169–190.

Schaeffer, M., & Carl, M. (2014). *Measuring the cognitive effort of literal translation processes.* Paper presented at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden.

Schilperoord, J. (1996). *It's about time: Temporal aspects of cognitive processes in text production*. Amsterdam: Rodopi.

Schleiermacher, F. (2012). On the different methods of translating (S. Bernofsky, Trans.). In L. Venuti (Ed.), *The translation studies reader* (3rd ed., pp. 43–63). London: Routledge.

Sharmin, S., Špakov, O., Räihä, K.-J., & Jakobsen, A. L. (2008). Where on the screen do translation students look while translating, and for how long? In S. Göpferich, A. L. Jakobsen, & I. M. Mees (Eds.), *Looking at eyes: Eye-tracking studies of reading and translation processing* (pp. 31–51). Copenhagen: Samfundslitteratur.

Shreve, G. M. (2002). Knowing translation: Cognitive and experiential aspects of translation expertise from the perspective of expertise studies. In A. Ruiccardi (Ed.), *Translation studies: Perspectives on an emerging discipline* (pp. 150–173). Cambridge: Cambridge University Press.

Shreve, G. M., & Angelone, E. (Eds.). (2010). *Translation and cognition*. Amsterdam: John Benjamins.

Silveira, F. d. S. D. d. (2011). *Working memory capacity and lexical access in advanced students of L2 English.* PhD thesis. Universidade Federal do Rio Grande do Sul, Brazil. Retrieved from http://www.lume.ufrgs.br/bitstream/handle/10183/39423/000824076.pdf?sequence=1

Sirén, S., & Hakkarainen, K. (2002). Expertise in translation. *Across Languages and Cultures, 3*(1), 71–82.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas* (pp. 223–231). Cambridge, MA.

Sousa, S. C. M. d., Aziz, W. F., & Specia, L. (2011). Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Proceedings of the International Conference of Recent Advances in Natural Language Processing* (pp. 97–103). Bulgaria.

Specia, L. (2011). *Exploiting objective annotations for measuring translation post-editing effort.* Paper presented at the 15th Conference of the European Association for Machine Translation, Leuven.

Specia, L., Raj, D., & Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation, 24*(1), 39–50.

Stymne, S. (2011). *Blast: A tool for error analysis of machine translation output.* Paper presented at the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon.

Sun, S. (2011). Think-aloud-based translation process research: Some methodological considerations. *Meta, 56*(4), 928–951.

Sun, S. (2015). Measuring translation difficulty: Theoretical and methodological considerations. *Across Languages and Cultures, 16*(1), 29–54.

Sun, S., & Shreve, G. M. (2014). Measuring translation difficulty: An empirical study. *Target, 26*(1), 98–127.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

Tatler, B. W., Kirtley, C., Macdonald, R. G., Mitchell, K. M., & Savage, S. W. (2014). The active eye: Perspectives on eye movement research. In M. Horsley, M. Eliot, B. A. Knight, & R. Reilly (Eds.), *Current trends in eye tracking research* (pp. 3–16). London: Springer.

TAUS. (2010). *MT post-editing guidelines*. Retrieved March 1, 2015, from https://www.taus.net/think-tank/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines

TAUS. (2014). Post-editing: Championing MT. Retrieved March 1, 2015 from https://postedit.taus.net/

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1927). *The measurement of intelligence*. New York: Bureau of Publications, Columbia University.

Tirkkonen-Condit, S. (2005). The monitor model revisited: Evidence from process research. *Meta, 50*(2), 405–414.

Tokowicz, N., Kroll, J. F., De Groot, A. M. B., & Van Hell, J. G. (2002). Number-of-translation norms for Dutch – English translation pairs: A new tool for examining language production. *Behavior Research Methods, Instruments, & Computers, 34*(3), 435–451.

Tomporowski, P. D. (2003). Performance and perceptions of workload among young and older adults: Effects of practice during cognitively demanding tasks. *Educational Gerontology, 29*(5), 447–466.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*(2), 127–154.

Vasconcellos, M. (1987). A comparison of MT post-editing and traditional revision. In K. Kummer (Ed.), *Proceedings of the 28th annual conference of the American Translators Association* (pp. 409-416). Medford: Learned Information.

Vidulich, M. A., & Tsang, P. S. (2012). Mental workload and situation awareness. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (4th ed., pp. 243–273). Hoboken: Wiley.

Vieira, L. N. (2014). Indices of cognitive effort in machine translation post-editing. *Machine Translation, 28*(3-4), 187–216.

Vonk, W., & Cozijn, R. (2003). On the treatment of saccades and regressions in eye movement measures of reading time. In J. Hyona, R. Radach, & H. deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 291–312). London: Elsevier.

Wierwille, W. W., & Williges, B. H. (1980). *An annotated bibliography on operator mental workload assessment* (Naval Air Test Center Report No. SY-27R-80). Patuxent River: Naval Air Test Center, System Engineering Test Directorate.

Wilson, R. A., & Keil, F. C. (Eds.). (1999). *The MIT encyclopedia of the cognitive sciences*. Cambridge: MIT Press.

Wilss, W. (1982). *The science of translation: Problems and methods*. Tübingen: Gunter Narr.

Wisniewski, G., Kübler, N., & Yvon, F. (2014). *A corpus of machine translation errors extracted from translation students exercises*. Paper presented at the International Conference on Language Resources and Evaluation (*LREC)*, Iceland. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1115_Paper.pdf

Woodrow, H. (1936). The measurement of difficulty. *Psychological Review, 43*(4), 341–365.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Boston: Houghton Mifflin.