Measuring the difficulty of text translation:

The combination of text-focused and translator-oriented approaches*

Yanmei Liu[1], Binghan Zheng[2] and Hao Zhou[2]

[1]Shandong University of Finance and Economics, China

[2] Durham University, U.K.

Abstract: This paper explores the impact of text complexity on translators' subjective perception of translation difficulty and on their cognitive load. Twenty-six MA translation students from a UK university were asked to translate three English texts with different complexity into Chinese, with their eye movements being recorded by an eye-tracker, and their cognitive load being self-assessed with a Likert scale before translation and NASA-TLX scales after translation. The results show that: (i) the intrinsic complexity measured by readability, word frequency and non-literalness was in line with the results received from informants' subjective assessment of translation difficulty; (ii) moderate and positive correlations existed between most items in the self-assessments and the indicator (fixation and saccade durations) obtained by the eye-tracking measurements; and (iii) the informants' cognitive load as indicated by fixation and saccade durations (but not for pupil size) increased significantly in two of the tree texts with the raise of source text complexity.

Keywords: text complexity; translation difficulty; cognitive load; eye-tracking; self-assessment

* This research is supported by Social Science Planning Program of Shandong Province (14CWXJ29) and Durham University Seedcorn Research Funding (04.14.290201)

1. Introduction

The significance of measuring the difficulty of a source text for translation pedagogy and research has received some attention in the past two decades (see, e.g., Hale and Campbell 2002; Jensen[1] 2009; Mishra et al. 2013; Sun and Shreve 2014). To investigate the degree of translation difficulty caused by the variable *text complexity,* researchers have based their examinations either on readability alone (Pavlović and Jensen 2009), or on a combination of readability and other indicators, such as word frequency, sentence structure and non-literalness (Sharmin et al. 2008; Jensen 2009). Measurement has generally been centred around the level of text complexity– for instance, character length, syllable length and sentence length – while ignoring other important factors, such as conceptual complexity, text organisation, or reader's background knowledge (Liu and Chiu 2011, 149). Nevertheless, the textual factors can account only partially for the text's level of translation difficulty (Sun and Shreve 2014, 98), since the construct of translation difficulty originates from the interaction between task and its translator. Therefore, translation difficulty should be measured on both texts and the profiling of translators who are working with the texts.

It was hypothesised that more complex texts would impose a heavier load on translators than easy ones, but it was uncertain to what extent the quantitative text measurement of intrinsic complexity would correlate with the informants' subjective measurement of their cognitive load and with a physiological measurement of cognitive effort.[2] Hvelplund (2011) and Sun and Shreve (2014) are among the few researchers who have adopted multiple measures, pertaining to both texts and translators. Hvelplund (2011) examined the effect of text complexity, measured by readability, word frequency and non-literalness, on translators' cognitive load as indicated by data on pupil size obtained from eye-tracking. He found no significant

differences in the means of pupil size among texts of different degrees of complexity. This result casts doubt on both the applicability of the above-mentioned three factors as a quantitative text measurement of intrinsic complexity, as well as on the reliability of a single pupil size as the physiological measurement. It was difficult to find a plausible explanation for the influence of text complexity on translators' cognitive load, due to the lack of informants' subjective assessment of translation difficulty. Sun and Shreve (2014) claim that NASA-TLX is a reliable subjective measurement for assessing translation difficulty, while text readability alone is only weakly correlated with the level of translation difficulty. [3]

Our research aims to integrate the indicators from the above two studies: the measurements of text complexity used in Hvelplund (2011) and the subjective assessments applied by Sun and Shreve (2014). In addition, eye movement data were collected to study the cognitive effort made by the informants when carrying out the translation tasks. According to Sun (2015) and Akbari and Segers (2017), measuring translation difficulty should incorporate manifold ways such as measuring text complexity, evaluating products and measuring translators' cognitive load. This study adopted measurements of text complexity and translators' cognitive load, with the aim of addressing the following research questions: (1) Are the indicators (cf. Hvelplund 2011) of quantitative text measurement of intrinsic complexity consistent with subjective self-assessments of translation difficulty? (2) What are the main differences in translators' cognitive load resulting from texts with different levels of translation difficulty? (3) Are the cognitive load levels as subjectively measured through the NASA-TLX questionnaire consistent with the physiological indicators supplied by eye-tracking data?

2. Measuring cognitive load

Empirical approaches to text difficulty may be traced back to Campbell (1999); before that it was mainly a subject of debate in the literature on reading research (Hale and Campbell 2002, 14). Translation difficulty can be operationalised by the load imposed on the performer's cognitive system and the effort invested in the execution of the task. Two effective techniques to measure cognitive load have been identified in previous research: *subjective* indices (rating scales) and *psycho-physiological* indices (e.g., pupil diameter, heart rate variability, event-related brain potentials; Paas and van Merriënboer 1994a, 357). The subjective indices in the present research included pre- and post-translation rating, described in detail in Section 4.2.

In this research, eye-tracking data were used as physiological indices indicating cognitive load, as in reading research. The results of studies using eye-tracking to assess reading task difficulty suggest that readers fixate longer when they are accessing long words (Just and Carpenter 1980; Rayner et al. 1996), low-frequency words (Just and Carpenter 1980; Inhoff 1984; Rayner and Fischer 1996; Rayner and Raney, 1996), novel (unfamiliar) words (Chafin, Morris and Seely 2001; Williams and Morris 2004), ambiguous words (Rayner and Duffy 1986; Sereno, O'Donnell, and Rayner 2006), and words that are not constrained by or predictable from the context (Ehrlich and Rayner 1981; Zola 1984; Rayner and Well 1996, Ashby et al. 2005). In addition to lexical factors, syntactic and discourse factors also influence the fixation duration (Staub and Rayner 2007). Readers spend more time integrating information from important clauses and making inferences at the ends of sentences (Just and Carpenter 1980, 329), spend more time processing metonymic referential descriptions and metaphorical expressions than literal expressions when they are at the beginning of a target sentence (Gibbs 1990), and also more time on reading garden path

sentences (Schotter and Rayner 2012, 91) than on conventional expressions. Moreover, structurally incoherent text segments attract more visual attention than coherent text segments (Vauras, Hyönä and Niemi 1992).

The findings from the above studies all point to fixation duration increasing when more complex information is being processed. Saccade duration should be taken into account as well (Irwin 2004, 128), because cognitive processing sometimes takes place during saccades. Mishra et al. (2013) used the sum of fixation and saccadic durations (henceforth FSD) as the processing time to measure translation difficulty index (TDI) for a sentence, and established that TDI is correlated with three properties of the input sentence, namely, length, degree of polysemy and structural complexity. This measurement was adopted in the present study as one of the two indicators of cognitive load measured by eye tracking.

In addition to FSD, measurements of pupil size or dilation are often used as indicators of the workload placed on a reader's cognitive system (Hvelplund 2014, 214). The positive correlation between pupil size and task difficulty was first suggested by Hess and Polt (1964). They found that, when simple multiplication problems were solved, an increase in task complexity elicited a strong pupillary response. In reading experiments, Just and Carpenter (1993) reported that more complex sentences required longer processing times and also yielded larger pupil dilations. Hyönä et al. (1995) compared simultaneous interpretation with other language processing tasks, and reported that the informants' average pupil sizes were quite different when performing tasks of various levels of difficulty. In written translation, Pavlović and Jensen (2009) found larger pupil sizes during TT reformulation than during ST comprehension. All these studies reached the same conclusion: the more difficult the task, the more dilated the pupils become.

Using the same experimental materials employed by Hvelplund (2011), the aim of the present study was to revisit the relationship between text complexity and translators' cognitive load.

3. Research Design

3.1 Informants

Twenty-six MA translation students (24 females and 2 males) from a UK university were recruited as informants on a voluntary basis.[4] They were considered representative of advanced learners of English-Chinese translation. After a pilot study and informants screening, a total of 22 informants were selected, with an average age of 23.78 years (range 22-24, *SD*=1.12 years). They were all native Mandarin Chinese speakers with English as their second language. None of them had been brought up in a bilingual context. Having learned English from the average age of 9.35 years (range 9-10, *SD*=0.43), these late bilinguals were ranked as highly proficient in English, with a mean IELTS[5] score at 7.42 (range 7-8, *SD*=0.35). They were all touch-typists and had normal or corrected-to-normal vision. To minimise negative influences on data quality, the informants were asked not to drink coffee or alcoholic beverages before the experiment, and female informants were asked not to wear heavy mascara. They were explained that anonymity and confidentiality would be ensured and they all signed a consent form before each experiment. Each informant received a £12 Tesco voucher as a reward for their work. The experiment was approved by the research ethics committee of the University.

3.2 Materials

Materials included a warm-up text and three experimental texts A, B, and C (Appendix I). The experimental texts were borrowed from Hvelplund (2011) with his permission and remained unchanged. They are all online newspaper articles with a general readership, which require no specialised knowledge for the purposes of translation. The texts are of similar length in terms of total number of characters and headlines. Three factors –readability, word frequency and non-literalness– served as indicators of text complexity. The linear progression from Text A over Text B to Text C in the level of reading difficulty, low frequency words and the number for non-literalness indicated that Text C was the most complex, Text A the least complex and Text B somewhere in between (see Figure 1).

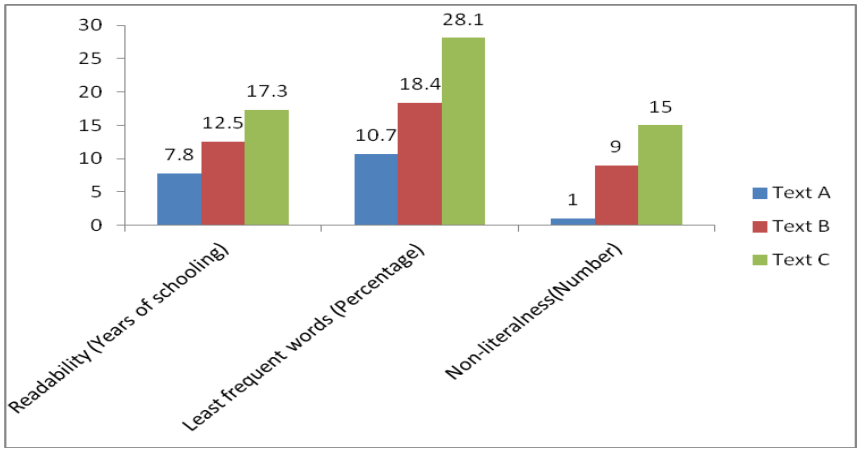


Figure 1. Summary of source text complexity by three indicators

3.3 Experimental settings

The experimental room was equipped with one stable light source on the ceiling in order to minimise the impact of light on the eyes. All the informants' eye movements were registered using a Tobii TX300 eye-tracker (300 Hz). The eye-tracker was connected to a 23" LCD monitor that was the presentation screen. The screen resolution was set at 1280*1024 pixels and the fixation radius was the default setting

of the Tobii system, 35 pixels per inch. The English source texts were displayed in the upper window of the key-logger Translog II user interface, [6] with a typeface New Times Roman at 20 point size, and double line spacing. The Chinese target texts were produced in the lower window, with the typeface SimSun with a 20 point size, and double line spacing. The I-VT Filter was applied as a fixation filter, which fixed the minimum fixation duration at 60 ms and the velocity threshold at 30 degrees/second.

3.4 Experimental procedure

The informants were tested individually in the university's eye-tracking laboratory. They were asked first to read three texts on paper, ordered in a Latin square design, and to rate the translation difficulty on a 0-10 Likert rating scale, with 0 as *extremely easy* and 10 as *extremely difficult*. They were given three minutes for the rating task, in order to avoid excessive processing of the texts before the eye-tracking experiment. Then, the informants were asked to sit approximately 60 cm away from the monitor; this was followed by a five-point calibration and validation procedure. After the acceptable calibration had been saved, each informant started to translate the warm-up text and then three experimental texts (in the same order as the initial rating tasks) with no time constraint. No online or offline aids or resources were provided during the experiment. The informants were allowed to take a break between tasks if requested. Finally, informants were asked to assess the cognitive load of their translation tasks based on the revised NASA-TLX scale applied by Sun and Shreve (2014) (see Appendix II). The complete session for each informant lasted roughly one hour.

4. Results

4.1 Quality assessment of eye-tracking data

The quality of collected eye-tracking data was assessed prior to data analysis. With reference to Hvelplund (2011), in the present research the following three criteria were adopted for the assessment: Gaze Time on Screen (GTS), Gaze sample to Fixation Percentage (GFP) and Mean Fixation Duration (MFD).

GTS indicates the amount of gaze time on the computer screen as a percentage of the total production time (Hvelplund 2011, 104). It was counted as [(fixations + saccades)/total production time)*100%]. Saccades were counted because fixations alone underestimate the duration of cognitive processing which still occurs during saccades (Irwin, 2004, p. 126). GTS scores lower than 73.1% (one SD below the mean) were considered invalid in our assessment.

GFP reveals the allocation of fixations and saccades in the gaze activity. It was calculated as [fixations/(fixations + saccades)*100%]. According to Hvelplund (2011), a saccade percentage higher than 15% indicates that some of the gaze sample rows reflect noise in the eye-tracking data. In line with Hvelplund's suggestion, this study set a GFP of 85.2% (one SD below the mean) as the threshold of valid data.

Mean Fixation Duration [total fixation duration/the number of fixations] proposed by Rayner (1998) has also been frequently used for assessing the quality of eye-tracking data. In this study, MFD lower than 241.60 ms (one SD below the mean) was considered as invalid data.

The data that met the requirements of at least two out of the above three criteria were included for further analysis (cf. Hvelplund 2011). Table 1 shows that the data from two informants (I7, I10) were deemed invalid and all their recordings were removed from further analysis. The percentage of invalid data was thus 8.33%.

Table 1. Summary of eye-tracking quality assessment with invalid data (marked as ×)

Text	Text A			Text B			Text C		
Criteria / Informant (I)	GTS	GFP	MFD	GTS	GFP	MFD	GTS	GFP	MFD
I3				×					
I6			×			×			
I7	×	×	×	×	×	×	×	×	×
I8			×			×			×
I9									×
I10	×	×	×					×	
I16			×			×			
I18	×			×					
I24			×						

4.2 Subjective measurements

4.2.1 Pre-translation rating

As mentioned in Section 3.2, the indicators of readability, word frequency and non-literalness suggested a progressive increase in complexity from Text A to B, and from B to C. All informants rated the pre-translation difficulty; Table 2 presents the statistical results.

Table 2. Statistical results of pre-translation rating of translation difficulty

Text	N	Mean	Sd.	Min	Max	Kendall's W	Chi-Square	Df	Sig.
A	22	4.00	1.10	1.50	6.00	.699	30.775	2	.000
B	22	4.45	1.28	2.00	7.00				
C	22	6.11	.72	5.00	7.50				

Table 2 shows that the translation difficulty score for Text A was slightly lower than that for Text B, and much lower than that for Text C. The mean pre-translation rating scores of Texts A, B and C were 4, 4.45 and 6.11 respectively, showing that, for the informants, their translation difficulty increased progressively, which is in line with the results of the text complexity test.

To further assess inter-rater reliability, that is, how well these informants agreed with each other on the levels of translation difficulty, Kendall's coefficient concordance was computed, with Kendall's W=0.699 and $p<0.01$, indicating a cut-off

point for denoting a strong agreement among the informants.[7] This result supports Sun and Shreve's (2014) finding that translators' pre-rating scores can to some extent predict the difficulty level of a translation.

4.2.2 Post-translation rating

The post-translation rating of the level of difficulty of the translation included four out of six NASA-TLX subscales: *Mental Demand*; *Effort*; *Frustration*; and *Performance*. As with the pre-translation rating, the informants were asked to rate the four subscales based on a 0-10 Likert scale; Table 3 presents the statistical results.

Table 3. Statistical results of post-translation rating of translation difficulty

Subscale	Text	Mean	Min	Max	Kendall's W	Chi-Square	Sig.
Mental Demand	A	4.18	1.5	6.5	.736	32.386	.000
	B	4.91	3	7			
	C	6.36	4	9			
Effort	A	4.39	1.5	6.5	.541	23.792	.000
	B	4.86	3	7			
	C	6.34	4	9			
Frustration	A	3.75	1.5	6	.681	29.949	.000
	B	4.39	1.5	6			
	C	6.09	3.5	9			
Performance	A	3.82	1	6.5	.342	15.027	.001
	B	4.39	3	6			
	C	5.3	4	7			

In Table 3, the mean, the minimum and the maximum values generally present a rising tendency from Text A to C, which implies that the informants accurately perceived differences in translation difficulty levels between the three texts. The results of Kendall's coefficient concordance on these four subscales (Kendall's W=0.736, $p<0.01$) show that the informants highly agreed that Text C was the most difficult text, while Text A was the least difficult. With regard to the subscales for Effort (Kendall's W=0.541, $p<0.01$) and Frustration (Kendall's W=0.681, $p<0.01$),

they moderately agreed that they had put the greatest amount of effort into translating Text C, and encountered maximal frustration during that translation, followed by Text B and then A. Of the four subscales in NASA-TLX, the Own Performance ratings produced a relatively low level of agreement among the informants (Kendall' W=0.342, $p<0.01$). The creator of NASA-TLX, Hart and Staveland (1988) also identified that Own Performance ratings were "relatively independent of the other ratings" (165). As a result, our first question can be answered positively, in that both pre- and post-translation ratings were consistent with the quantitative text measurements of intrinsic complexity.

4.3 Physiological measurements

To explore the changes in the informants' cognitive load when translating texts of different complexity levels, physiological data, including FSD and pupil dilation, were also analysed in this study.

4.3.1 Fixation and saccade duration (FSD)

As can be seen from Figure 2, the means of the sum of fixation and saccadic durations, or FSD, display a rising tendency from Text A to Text C, and the increase from Text A to Text B is smaller than that from Text B to Text C. However, the mean FSD of each informant in the three texts displays an intertwined tendency (see Figure 3).

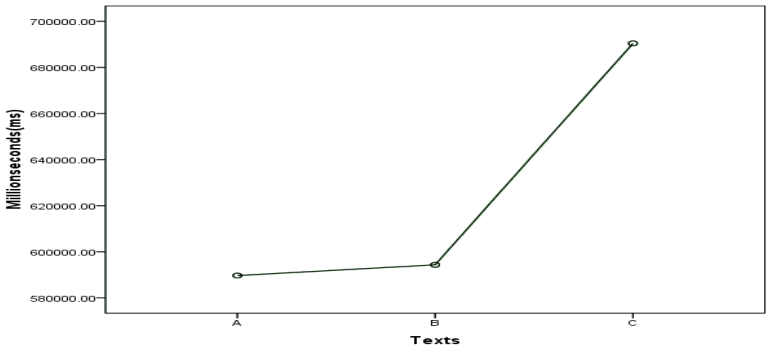


Figure 2. Mean FSD in Texts A, B and C

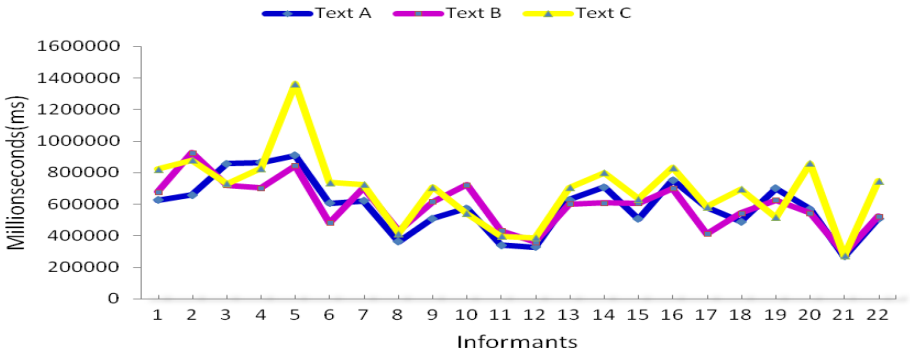


Figure 3. Mean FSD for each informant translating Texts A, B and C

Table 4. Normality test of mean FSD

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Texts	Statistic	Df	Sig.	Statistic	Df	Sig.
FSD	A	.099	22	.200*	.968	22	.670
	B	.112	22	.200*	.981	22	.936
	C	.157	22	.171	.914	22	.058

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

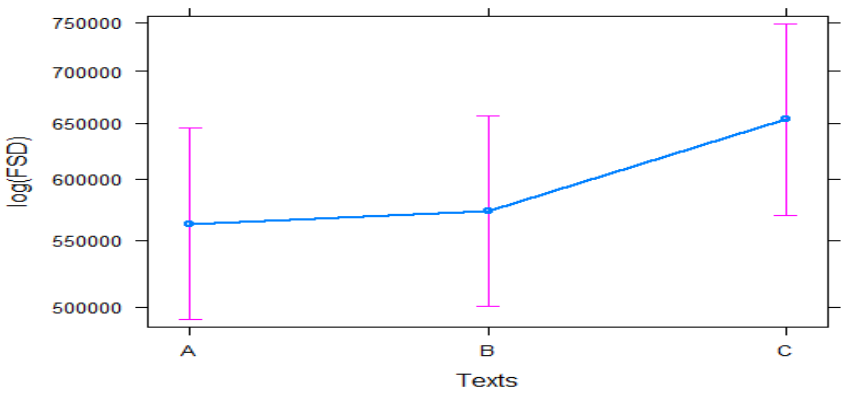


Figure 4. The effect of text complexity on FSD

Thus, an additional statistical analysis was performed to determine whether the observed differences in FSD among the three texts were statistically significant. Table 4 shows that the data of the three groups were normally distributed (both KS and SW tests have $p>0.05$). Based on this result, linear mixed-effects regression (LMER) model was conducted with text complexity as the fixed effect and informants as the random effect. The result (see Figure 4) showed that FSD are significantly different between Text A and C ($t=3.659$, $p=0.001$), Text B and C ($t= 3.211$, $p=0.003$) but not between Text A and B ($t=0.447$, $p=0.66$).

4.3.2 Pupil dilation

The means of pupil sizes for Texts A (2.99), B (2.99) and C (2.88) were very close to each other. In order to explore the potentially concealed variance among pupil sizes in the three texts, a one-way ANOVA test was conducted based on the results of the normal distribution test and the homogeneity test of variance. The result of KS normal distribution test (see Table 5) shows that the data of the three groups were normally distributed (the KS Z=0.852 for text A, Z=0.794 for text B and Z=0.682 for text C, $p>0.05$). The result of the homogeneity test of variance (see Table 6) shows that the group variances were homogeneous ($p>0.05$). The result of the one-way ANOVA test (Table 7) shows that there were no statistically significant differences among the three texts with regard to means of pupil size ($F=0.009$, $p>0.05$).

Table 5. Normality test of pupil size statistics in the three texts

		Text A	Text B	Text C
N		22	22	22
Normal Parameters[a]	Mean	2.992	2.986	2.981
	Std. Deviation	.277	.270	.267
Most Extreme Differences	Absolute	.182	.169	.145
	Positive	.182	.165	.145
	Negative	-.156	-.169	-.131
Kolmogorov-Smirnov Z		.852	.794	.682
Sig. (2-tailed)		.462	.554	.741

a. Test distribution is normal.

Table 6. Test of homogeneity of variances (pupil size)

Levene Statistic	Df1	Df2	Sig.
.007	2	63	.993

Table 7. ANOVA test of pupil size among Texts A, B and C

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.001	2	.001	.009	.991
Within Groups	4.650	63	.074		
Total	4.651	65			

Although our results are largely in line with Hvelplund's (2011), we did not expect that Text C (the most complex text) would have the lowest mean value of pupil sizes. To determine whether the 'acclimatisation effect' (Hyönä et al. 1995; O'Brien 2006) was operating in this study, the pupil size data were classified into three groups according to the order of texts translated. The means of pupil sizes were regrouped according to the translating sequence (A-B-C, B-C-A and C-A-B). Figure 5 reveals that the first text in each sequence always induced the largest mean value of pupil size. Friedman tests indicated that pupil sizes were significantly influenced by task sequences (p=0.028), but not by text complexity (p=0.483).

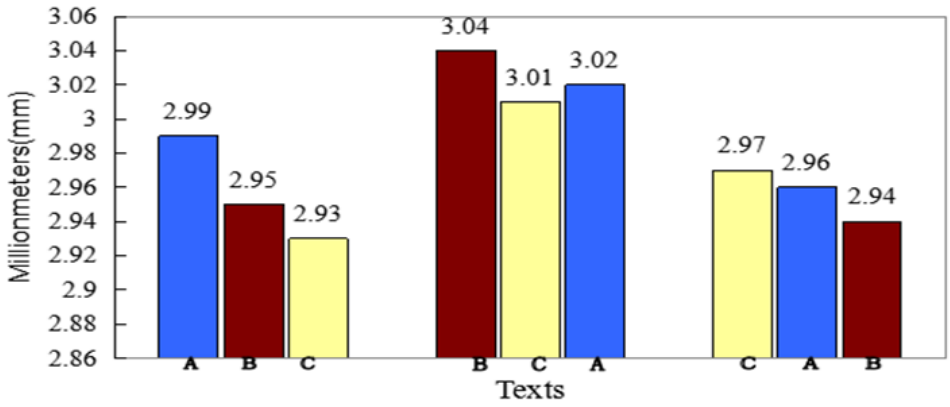


Figure 5. Means of pupil size in different task sequences

4.4 The link between subjective and physiological measurements

This section presents the results of studying the relationship of the informants' perceptions of translation difficulty with their cognitive load as indicated by eye-tracking data. We assumed that post-translation ratings would reflect the informants' perceptions more accurately than pre-translation ratings, since having actually translated the texts might have resulted in an improved ability to accurately evaluate the difficulty levels of those texts. Thus, the data of the post-translation ratings, including those related to *Mental Demand*, *Effort*, *Frustration* and *Own Performance*, were used for further statistical analysis.

Table 8. Normality test of post-translation rating scores

		Mental Demand	Effort	Frustration	Own Performance
N		66	66	66	66
Normal Parameters[a]	Mean	5.152	5.197	4.742	4.500
	Std. Deviation	1.468	1.422	1.574	1.237
Most Extreme Differences	Absolute	.156	.138	.132	.157
	Positive	.132	.133	.132	.157
	Negative	-.156	-.138	-.126	-.146
Kolmogorov-Smirnov Z		1.266	1.122	1.072	1.275
Sig. (2-tailed)		.081	.161	.200	.077

a. Test distribution is normal.

The normal distribution test (see Table 8) revealed that there was no violation of the normality assumption in any of the four subscales (p>0.05) of the post-translation rating. The scatterplot showed that there was a positive linear relationship between NASA-TLX measurements (the average of four subscales) and FSD in all three texts (see Figure 6).

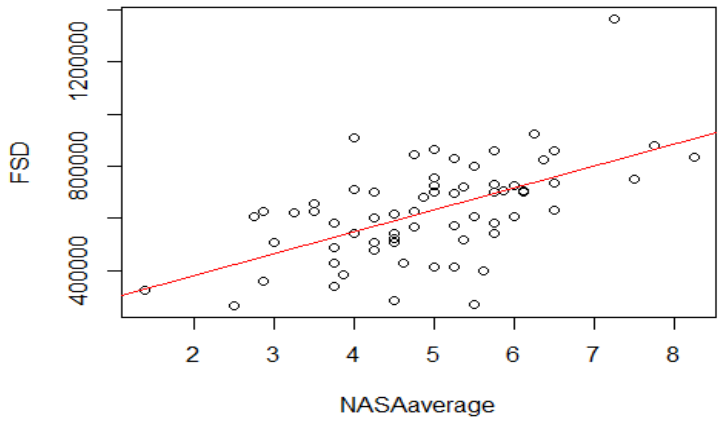


Figure 6. The correlation between NASA-TLX measurements (in average) and FSD

As the data from NASA-TLX and FSD met the following three requirements namely, ratio data, linear relationship and normal distribution, we run a Pearson's correlation coefficient test to measure the relationship between subjective assessment and eye-tracking data. Most of the values of Pearson's r in these tests (see Table 9) were above 0.40, with the exception of one value that equaled 0.389 (the correlation coefficient between Performance and FSD in Text C). Most of the positive correlations between the four subscales of NASA-TLX and the eye-tracking data were significant (p<0.05), apart from three p-values (in bold) that were slightly greater than 0.05. In short, almost all the subscales of the subjective assessments had a moderately or strong positive correlation with the eye-tracking data indicated by FSD.

Table 9. Pearson correlation test between subjective measures and FSD[8]

Subscale	Pearson Correlation	Text A	Text B	Text C
Mental Demand	Pearson's r	.582	.499	.584
	Sig. (2-tailed)	.004	.018	.004
Effort	Pearson's r	.529	.605	.642
	Sig. (2-tailed)	.011	.003	.001
Frustration	Pearson's r	.564	.417	.443
	Sig. (2-tailed)	.006	**.054**	.039
Performance	Pearson's r	.474	.415	.389
	Sig. (2-tailed)	.026	**.055**	**.073**

5. Discussion

This study aimed to investigate whether texts of different complexity, as indicated by readability, word frequency and non-literalness, had a correlation with translators' perceptions of translation difficulty and, accordingly, might induce a different amount of cognitive effort. The first and the third research questions can have a clear answer. The quantitative text measurements of intrinsic complexity were consistent with the informants' self-assessments of translation difficulty. In addition, there was a moderate and positive correlation between cognitive load as measured by NASA-TLX and cognitive load as indicated by FSD. The diversified results obtained from the physiological measurement make the answer to the second question much more complicated. Variations of cognitive load on account of text complexity were confirmed by the FSD data. Nevertheless, the pupil size does not yield such corresponding variances. The three types of measurement are discussed separately below.

5.1 Quantitative text measurements

Quantitative measurements of intrinsic difficulty focus on the linguistic properties of texts, which refer to word frequency, readability and non-literalness in this article. Despite the small correlation coefficient between the single indicator of readability and translation difficulty level (Sun and Shreve 2014), the validity of these three

properties as indicators of text complexity has been verified to some extent. For instance, they clearly correlated with informants' judgements, as shown in Section 4.2. The cognitive load indicated by FSD traced a rising tendency with the increase in text complexity. Separately, the individual component of these three elements quantifies text difficulty in only one aspect located at a certain place in the text, while the combination of three elements reflects the interactive influence on text difficulty in the whole, as Carpenter and Just (1989, 61) stated that "readability is not just a function of the difficulty of a given portion of text, but is also a function of how that difficulty impinges on the maintenance of other information". Thus the level of task complexity depends much on the number of elements to be processed simultaneously and on the degree of elements interactivity. In cognitive load theory, element interactivity has been used as the basic, defining mechanism of intrinsic cognitive load. The higher the number of interacting elements, the heavier the working memory load (Sweller 2010, 123-124). The combination of three elements as well as the linear progression in the difficulty level of them prove more effective for judging text complexity and easier to be perceived by the informants.

However, the data of pupil size offered no strong evidence supporting the notion that increased text complexity requires greater cognitive load. This result is consistent with Hvelplund (2011), and might be attributed either to the intrinsic defects of cognitive indicators, or to the distribution of the translators' cognitive resources, which is explained in more detail in Section 5.2.

5.2 Physiological measurements

The increased text complexity indeed costs a greater cognitive effort, as revealed by FSD. The reason for this result might be induced by the interactive influence of

linguistic features on text complexity. Judged from one single factor, for instance; non-literalness, there is a wider gap of complexity between Text A and the other two texts (see Figure 1). However, the combining effects of this factor with the other two linguistic features (readability and word frequency) contributed differently on the degree of cognitive load which could be seen by the informants' pre-translation rating scores (see Table 2). The perceived cognitive load from Text C (Mean=6.11) is much higher than that from Text A (Mean=4.00) and B (Mean=4.45), while the difference of cognitive load between Text A and B is small. This is consistent with the informants' cognitive effort indicated by FSD.

Besides, we tried to minimise the adverse effects of external factors by asking the informants to translate with no access to the auxiliary instruments. Under this condition, the informants' attention was concentrated on the comprehension and transformation of the texts, rather than on searching external resources and selecting potentially optimal solutions to translation problems. Consequently, the cognitive effort undeviatingly reflects informants' response to the cognitive load imposed by the texts. This explains why FSD is significantly longer in Text C than in Text A and B, but has no significant difference between Text A and B.

On the other hand, the variation in cognitive load was not observable in the data on pupil size. According to Iqbal et al. (2004) and Schultheis and Jameson (2004), pupil dilation may not always be sensitive to the variation of task load. As a matter of fact, the results of both this study and that of Hvelplund (2011) suggest that pupil size may not be a suitable indicator of cognitive load for translation tasks lasting for a relatively long period, for the following the reasons:

Firstly, pupil dilation may be influenced by a variety of factors, such as ambient illumination, task complexity, gaze angle and coffee/alcohol intake. Some factors

were controlled – e.g., the lighting was maintained at a constant level, as was the brightness of the screen, and no coffee/alcohol drinking was allowed – but it was still difficult to control some other influencing factors, such as the moving gaze angle. During the translation process, the gaze position may change slightly as a result of the informant's habitual and occasional looking at the keyboard, even if touch-typing is requested. Also, when informants move their eyes during experiments, their pupils may be at different angles to and distances from the monitoring camera of the eye tracker. This, in turn, leads to inconsistency in the measurement of pupil sizes. "This effect is especially strong if the camera is located below the eye" (Pomplun and Sunkara 2003, 542).

Secondly, this study did not examine the pupil dilation at particular points of difficulty, such as non-literal expressions. Thus, specific, relevant pupil dilations might be concealed in the mean values of the whole text. Some more difficult words or expressions were assumed to perhaps induce higher cognitive loads in the informants, but they would not lead to larger average pupil diameters throughout the whole text. This confirms Schultheis and Jameson (2004) assumption, who found that changes in pupil sizes corresponded to the level of cognitive load in the subtasks, but not in the whole tasks. They concluded that "pupil size may differ between easy and difficult conditions only in certain periods of a task" (234). In order to use pupil size as a cognitive load indicator, Schultheis and Jameson (2004, 227) proposed that at least three of the following five conditions need to apply: (a) constant lighting; (b) avoidance of eye movements; (c) use of nonvisual (e.g., acoustic) stimuli; (d) use of many similar, short tasks, and (e) evaluating only mean values averaged across tasks and subjects. Conditions (a) and (e) have been applied to our study, but not the rest. This may explain why we found no effect of text difficulty on pupil size.

5.3 Subjective judgements

The consistency of difficulty assessment we found between subjective judgements and quantitative text measurements, coupled with the positive correlation between self reports and physiological measurements, suggest that using rating scales for self reports could be a more reliable method for assessing translation difficulty. This result lends support to Paas (1992) and Paas and van Merriënboer (1994b), which claimed that self reports are "reliable, unobtrusive and more sensitive to relatively small differences in cognitive load" (Sweller et al. 1998, 268).

These findings, coupled with Sun and Shreve (2014), suggest that subjective judgements can be considered a valuable tool for estimating translation difficulty, in view of its easy accessibility and relatively high reliability. These merits, however, may be accompanied by the flaws in this method. "Individual differences such as previous experience, depth of background knowledge, and domain skills" (Liu and Chiu 2011, 152) may produce great discrepancies in perceptions of the difficulty level of a text. In addition to the unavoidable subjectivity resulting from personal capability for the task manipulations, personal predictions are sensitive to external factors such as working conditions and the translation brief (e.g., routine practice or customer demand). Some researchers argue that a "subjective feeling of difficulty is essentially dependent on the time pressure involved in performing the task" (Cain 2007, 8). Furthermore, personal assessments do not seem to tap on unconscious or automatic processes. In view of all these arguments, subjective judgements can be seen as the reliable overall measure (Johannsen 1979)，but it is only "a gross indicator of stress level and have little diagnostic value" (Cain 2007, 8).

6. Conclusion

In the hope of finding a convenient and effective way of measuring translation difficulty, this study designed a set of experiments to explore the interactions between text complexity and cognitive load, with a multiple comparison of subjective indicators (pre- and post-translation rating), quantitative text indicators (readability, word frequency and non-literalness) and physiological indicators (FSD, pupil dilation). The findings can be summarised as follows:

First, the validity of text features (readability, word frequency and non-literalness) as quantitative text indicators of intrinsic difficulty was confirmed by the subjective ratings. This result at least suggests the effectiveness of the reciprocal influence of more quantitative text indicators on the assessment of text complexity.

Second, the results suggest that subjective ratings based on NASA-TLX are more sensitive to comparable translation difficulty and cognitive load levels retrieved by the indicators of readability, word frequency and non-literalness. The informants' self-assessments of translation difficulty were consistent with quantitative measurements of text complexity. Furthermore, post-translation rating of *mental demand*, *effort*, *frustration* and *performance* had a slightly higher positive correlation with cognitive load as indicated by FSD. Thus, subjective judgements may still serve as a more cost-effective approach to evaluating the translation difficulty of a text, despite their flaws caused by subjectivity and their incapability of accounting for unconscious or automatic translation processes.

Third, the effect of text complexity on the translators' cognitive load was revealed by the indicator of FSD, but not by pupil size. The latter tends to be more susceptible to the order of text presentation than to the complexity of texts. More experimental evidence would be helpful to work out the association between

eye-tracking measurements and text complexity.

The results yielded in this study may contribute to establishing measurements for testing the difficulty of translation materials, and consequently enable translation teachers or assessors to set translation difficulty levels in translation pedagogy. However, we are mindful of some limitations existing in this study: such as the limited number of source texts, unified text type and domain, and the single group of student informants recruited. Future studies could diversify the design of task types and select informants with different professional levels. A comparative analysis on eye-tracking data between source and target texts might clarify to what extent translation difficulty is a comprehension or a production phenomenon. Furthermore, verbal protocols, Translog, and quality assessment data could be included in order to strengthen data triangulation.

Acknowledgements

We would like to thank Professor Ricardo Muñoz Martín, the anonymous reviewers and editors who provided constructive feedback that has helped strengthen this article.

Notes

1. Authors referred to as Jensen or Hvelplund in this text are one and the same person.
2. In this article, cognitive load refers to the demand of cognitive resources a task put on a translator, while cognitive effort is the actual amount of cognitive resources that translators put into task processing.
3. NASA –TLX (NASA Task Load Index) is a multidimensional scale developed by Hart and Staveland (1988) to measure subjective workload. Six workload-related subscales include: mental demand, physical demand, temporal demand, effort, performance and frustration level. Each subscale is presented as a line divided into 20 equal intervals anchored by bipolar descriptors (e.g., low/ high, good/poor).
4. It was anticipated that the gender imbalance would not have a decisive influence on the results of the study (Hvelplund, 2011).
5. The International English Language Testing System (IELTS) is one of the most widely accepted English language proficiency tests for higher education and global migration. It is reported as band scores on a scale from 1 (the lowest) to 9 (the highest).
6. Translog‖ was only used to display the source texts and input the target texts. Translog data were not analysed in this research project.
7. Kendall's W ranges from 0, *complete disagreement,* to 1, *perfect agreement.* The responses are regarded as very strong agreement if Kendall's W is between 0.91-1, strong agreement if Kendall's W is between 0.71-0.90, moderate agreement if

Kendall's W is between 0.51-0.70, weak agreement if Kendall's W is between 0.31-0.50, and lack of agreement if Kendall's W is 0.0-0.30 (LeBreton and Senter 2008, 836).

8. Coefficient values usually range from +1 through 0 to -1, with +1 indicating a perfect positive relationship, -1 a perfect negative relationship, and 0 no relationship. The strength of the correlation, according to Evans (1996) depends on the absolute value of r: 0.00-0.19, "very weak"; 0.20-0.39, "weak"; 0.40-0.59, "moderate"; 0.60-0.79, "strong"; 0.80-1.0, "very strong".

References

Akbari, Alireza, and Winibert Segers. 2017. “Translation Difficulty: How to Measure and What to Measure.” *Lebende Sprachen* 62 (1): 3-29.

Ashby, Jane, Keith Rayner, and Charles Clifton Jr. 2005. “Eye Movements of Highly Skilled and Average Readers: Differential Effects of Frequency and Predictability.” *The Quarterly Journal of Experimental Psychology Section A* 58 (6): 1065-1086.

Cain, Brad. 2007. *A Review of the Mental Workload Literature*. Technical Report, Defence Research and Development Canada Toronto.

Campbell, Stuart. 1999. “A Cognitive Approach to Source Text Difficulty in Translation.” *Target* 11 (1): 33-63.

Carpenter, Patricia A., and Marcel A. Just. 1989. “The Role of Working Memory in Language Comprehension.” In *Complex Information Processing: The Impact of Herbert A. Simon*, ed. by David Klahr, and Kenneth Kotovsk, 31-68. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Chaffin, Roger, Robin K. Morris, and Rachel E. Seely. 2001. “Learning New Word Meanings from Context: A Study of Eye Movements.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27(1): 225-235.

Ehrlich, Susan F., and Keith Rayner. 1981. “Contextual Effects on Word Perception and Eye Movements During Reading.” *Journal of Verbal Learning and Verbal Behavior* 20(6): 641-655.

Evans, James D. 1996. *Straightforward Statistics for the Behavioral Sciences*. Pacific Grove, CA: Brooks/Cole Publishing Co.

Gibbs Jr, Raymond W. 1990. “Comprehending Figurative Referential Descriptions.” *Journal of Experimental Psychology: Learning, Memory and Cognition* 16 (1): 56-66.

Hale, Sandra, and Stuart Campbell. 2002. “The Interaction between Text Difficulty and Translation Accuracy.” *Babel* 8 (1): 14-33.

Hart, Sandra G., and Lowell E. Staveland. 1988. “Development of NASA-TLX (Task

Load Index): Results of Empirical and Theoretical Research." In *Human Mental Workload*. ed. by Hancock, Peter. A. and Najmedin Meshkati, 139-183. Amsterdam: North-Holland.

Hess, Eckhard H., and James M. Polt. 1964. "Pupil Size in Relation to Mental Activity During Simple Problem-solving." *Science* 143 (3611): 1190-1192.

Hvelplund, Kristian Tangsgaard. 2011. *Allocation of Cognitive Resources in Ttranslation: An Eye-tracking and Key-logging Study*. Doctoral dissertation. Copenhagen Business School.

_____2014. "Eye Tracking and the Translation Process: Reflections on the Analysis and Interpretation of Eye-tracking Data." In *Minding Translation /Con la traducción en mente*, ed. by Muñoz Martín, Ricardo, 201–224. San Vicente del Raspeig: Publicaciones de la Universidad de Alicante.

Hyönä, Jukka, Jorma Tommola, and Anna-Mari Alaja. 1995. "Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks." *The Quarterly Journal of Experimental Psychology* 48 (3): 598-612.

Inhoff, Albrecht Werner. 1984. "Two Stages of Word Processing During Eye Fixations in the Reading of Prose." *Journal of Verbal Learning & Verbal Behavior* 23 (5): 612-624.

Iqbal, Shamsi T., Xianjun Sam Zheng, and Brian P. Bailey. 2004. "Task-evoked Pupillary Response to Mental Workload in Human-computer Interaction." In *CHI'04 Extended Abstracts on Human Factors in Computing Systems*, ed. by Dykstra-Erickson, Elizabeth, and Manfred Tscheligi, 1477-1480. Vienna, Austria.

Irwin, David E. 2004. "Fixation Location and Fixation Duration as Indices of Cognitive Processing." In *The Interface of Language, Vision, and Action: Eye Movements and Visual World*, ed. by Henderson, John, and Fernanda Ferreira, 105-134. New York: Psychology Press.

Jensen, Kristian T. H. 2009. "Indicators of Text Complexity." In *Behind the Mind: Methods, Models and Results in Translation Process Research*, ed. by Susanne Göpferich, Arnt Lykke Jakobsen, and Inger M. Mees, 61-80. Copenhagen: Samfundslitteratur.

Johannsen, Gunnar. 1979. “Workload and Workload Measurement.” In *Mental Workload: Its Theory and Measurement*, ed. by Neville Moray, 3-11. New York: Springer Science & Business Media.

Just, Marcel A., and Patricia A. Carpenter. 1980. “A Theory of Reading: From Eye Fixations to Comprehension.” *Psychological Review* 87 (4): 329-354.

_____ 1993. “The Intensity Dimension of Thought: Pupillometric Indices of Sentence Processing.” *Canadian Journal of Experimental Psychology* 47 (2): 310-339.

LeBreton, James M., and Jenell L. Senter. 2008. “Answers to 20 Questions about Interrater Reliability and Interrater Agreement.” *Organizational Research Methods* 11 (4): 815-852.

Liu, Minhua, and Yu-Hsien Chiu. 2011. “Assessing Source Material Difficulty for Consecutive Interpreting.” In *Interpreting Chinese, Interpreting China*, ed. by Robin Setton, 135-156. Amsterdam/Philadelphia: John Benjamins.

Mishra, Abhijit, Pushpak Bhattacharyya, and Michael Carl. 2013. “Automatically Predicting Sentence Translation Difficulty.” In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics(Volume 2: Short Papers)*, 346-351. Sofia: Bulgaria.

O’Brien, Sharon. 2006. “Eye-tracking and Translation Memory Matches.” *Perspectives* 14 (3): 185-205.

Paas, Fred 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, Fred G. W. C., and Jeroen J. G. van Merriënboer. 1994a. “Instructional Control of Cognitive Load in the Training of Complex Cognitive Tasks.” *Educational Psychology Review* 6 (4): 351-371.

Paas, Fred G. W. C., and Jeroen J. G. van Merriënboer. 1994b. “Variability of Worked Examples and Transfer of Geometrical Problem-solving Skills: A Cognitive-load Approach.” *Journal of Educational Psychology* 86 (1): 122-133.

Pavlović, Nataša, and Kristian Jensen. 2009. “Eye Tracking Translation

Directionality.” In *Translation Research Projects 2*, ed. by Anthony Pym and Alexander Perekrestenko, 93-109. Tarragona: Intercultural Studies Group.

Pomplun, Marc, and Sindhura Sunkara. 2003. “Pupil Dilation as an Indicator of Cognitive Workload in Human-computer Interaction.” In *Proceedings of the 10th International Conference on HCI* (Vol.3), 542-546. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Rayner, Keith. 1998. “Eye Movements in Reading and Information Processing: 20 Years of Research.” *Psychological Bulletin* 124 (3): 372-422.

Rayner, Keith, and Arnold D. Well. 1996. “Effects of Contextual Constraint on Eye Movements in Reading: A Further Examination.” *Psychonomic Bulletin and Review* 3 (4):504-509.

Rayner, Keith, and Gary E. Raney. 1996. “Eye Movement Control in Reading and Visual Search: Effects of Word Frequency.” *Psychonomic Bulletin & Review* 3 (2): 245-248.

Rayner, Keith, and Martin H. Fischer. 1996. “Mindless Reading Revisited: Eye Movements during Reading and Scanning Are Different.” *Perception & Psychophysics* 58 (5): 734-747.

Rayner, Keith, and Susan A. Duffy. 1986. “Lexical Complexity and Fixation Times in Reading: Effects of Word Frequency, Verb Complexity, and Lexical Ambiguity.” *Memory & Cognition* 14 (3):191–201.

Rayner, Keith, Sara. C. Sereno, and Gary E. Raney 1996. “Eye Movement Control in Reading: A Comparison of Two Types of Models.” *Journal of Experimental Psychology: Human Perception and Performance* 22 (5): 1188-1200.

Schotter, Elizabeth R., and Keith Rayner. 2012. “Eye Movements in Reading: Implications for Reading Subtitles.” In *Eye Tracking in Audiovisual Translation*, ed. by Perego Elisa, 83-104. Roma: Aracne Editrice.

Schultheis, Holger, and Anthony Jameson. 2004. “Assessing Cognitive Load in Adaptive Hypermedia Systems: Physiological and Behavioural Methods.” In *Adaptive Hypermedia and Adaptive Web-based Systems,* ed. by Wolfgang Nejdl and Paul. De Bra, *2*25-234. Berlin: Springer.

Sereno, Sara. C., Patric J. O'donnell, and Keith Rayner. 2006. "Eye Movements and Lexical Ambiguity Resolution: Investigating the Subordinate-bias Effect." *Journal of Experimental Psychology: Human Perception and Performance* 32 (2): 335-350.

Sharmin, Selina., et al. 2008. "Where on the Screen Do Translation Students Look While Translating, and for How Long?" In *Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing*, ed. by Jakobsen, Arnt Lykke, Susanne Göpferich, and Inger M. Mees, 31-51. Copenhagen: Samfundslitteratur.

Staub, Adrian, and Keith Rayner. 2007. "Eye Movements and On-Line Comprehension Processes." In *The Oxford Handbook of Psycholinguistics*, ed. by M. Gareth Gaskell and Gerry Altmann. 327-342. Oxford: Oxford University Press.

Sun, Sanjun. 2015. "Measuring Translation Difficulty: Theoretical and Methodological Considerations." *Across Languages and Cultures* 16 (1): 29-54.

Sun, Sanjun, and Gregory M. Shreve 2014. "Measuring Translation Difficulty: An Empirical Study." *Target* 26 (1): 98-127.

Sweller, John. 2010. "Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load." *Educational Psychology Review* 22 (2): 123-138.

Sweller, John, Jeroen J. G. van Merrienboer, and Fred G. W. C. Paas. 1998. "Cognitive Architecture and Instructional Design." *Educational Psychology Review* 10 (3): 251-296.

Vauras, Marja, Jukka Hyönä, and Pekka Niemi. 1992. "Comprehending Coherent and Incoherent Texts: Evidence from Eye Movement Patterns and Recall Performance." *Journal of Research in Reading* 15 (1): 39-54.

Williams, Rihana, and Robin Morris. 2004. "Eye Movements, Word Familiarity, and Vocabulary Acquisition." *European Journal of Cognitive Psychology* 16 (1-2): 312-339.

Zola, David. 1984. "Redundancy and Word Perception during Reading." *Perception & Psychophysics* 36 (3): 277-284.

Appendix I

Warm-up text

A wedding now costs $35,000

Source: *Daily Mail Online* (3 February 2017)

Study reveals tying the knot costs more than ever as couples look tomake their ceremony more lavish. The average wedding last year cost $35,329, it's been revealed. Record-breaking figure comes from The Knot 2016 Real Weddings Study. It's a jump from $32,641 - the average cost of a wedding in the 2015 study. The most expensive place to tie the knot is in Manhattan ($78,464) and the cheapest is in Arkansas ($19,522).

Number of characters with spaces: 415

Length of headline in characters with spaces: 27

Experimental texts (cf. Jensen 2009)

(Text A) Killer nurse receives four life sentences

Source: *The Independent* (4 March 2008)

Hospital nurse Colin Norris was imprisoned for life today for the killing of four of his patients. 32 year old Norris from Glasgow killed the four women in 2002 by giving them large amounts of sleeping medicine. Yesterday, he was found guilty of four counts of murder following a long trial. He was given four life sentences, one for each of the killings. He will have to serve at least 30 years. Police officer Chris Gregg said that Norris had been acting strangely around the hospital. Only the awareness of other hospital staff put a stop to him and to the killings. The police have learned that the motive for the killings was that Norris disliked working with old people. All of his victims were old weak women with heart problems. All of them could be considered a burden to hospital staff.

Number of characters with spaces: 837

Length of headline in characters with spaces: 41

(Text B) Families hit with increase in cost of living

Source: *The Times* on 12 February 2008

British families have to cough up an extra £1,300 a year as food and fuel prices soar at their fastest rate in 17 years. Prices in supermarkets have climbed at an alarming rate over the past year. Analysts have warned that prices will increase further still, making it hard for the Bank of England to cut interest rates as it struggles to keep inflation and the economy under control. To make matters worse, escalating prices are racing ahead of salary increases, especially those of nurses and other healthcare professionals, who have suffered from the government's insistence that those in the public sector have to receive below-inflation salary increases. In addition to fuel and food, electricity bills are also soaring. Five out of the six largest suppliers have increased their customers' bills.

Number of characters with spaces: 846
Length of headline in characters with spaces: 44

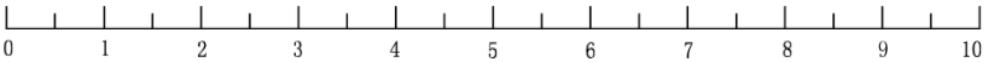
(Text C) Spielberg shows Beijing red card over Darfur
Source: *The Daily Telegraph* on 13 February 2008
In a gesture sure to rattle the Chinese Government, Steven Spielberg pulled out of the Beijing Olympics to protest against China's backing for Sudan's policy in Darfur. His withdrawal comes in the wake of fighting flaring up again in Darfur and is set to embarrass China, which has sought to halt the negative fallout from having close ties to the Sudanese government. China, which has extensive investments in the Sudanese oil industry, maintains close links with the Government, which includes one minister charged with crimes against humanity by the International Criminal Court in The Hague. Although emphasizing that Khartoum bears the bulk of the responsibility for these ongoing atrocities, Spielberg maintains that the international community, and particularly China, should do more to end the suffering.

Number of characters with spaces: 856
Length of headline in characters with spaces: 44

Appendix II
The Adapted NASA Task Load Index for Measuring Translation Difficulty (cf. Sun and Shreve 2014)

1.

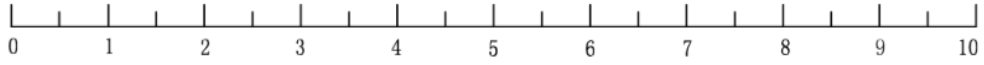
Mental Demand How mentally demanding was the task?



Low High

2.

Effort How hard did you have to work to accomplish your level of performance?



Low High

3.

Frustration How insecure, discouraged, irritated, stressed, and annoyed were you?

Low High

4.

Performance How successful were you in accomplishing what you were asked to do?

Poor Good

Address for correspondence

Binghan ZHENG (Corresponding author)
School of Modern Languages and Cultures, Durham University
Elvet Riverside, New Elvet, Durham, DH1 3JT, United Kingdom
Email: binghan.zheng@durham.ac.uk
https://orcid.org/0000-0001-5302-4709

Co-author information

Yanmei LIU
Shandong University of Finance and Economics
Email: lymcx@126.com

Hao ZHOU
Durham University
Email: hao.zhou@durham.ac.uk