Assessing source material difficulty for consecutive interpreting

Quantifiable measures and holistic judgment*

Minhua Liu and Yu-Hsien Chiu
Fu Jen University / National Taiwan Normal University

Motivated by the need for better control of standards of a certification examination for interpreters in Taiwan, this exploratory study aimed at identifying indicators that may be used to predict source material difficulty for consecutive interpreting. A combination of quantifiable measures — readability level, information density and new concept density — was used to examine different aspects of three English source materials. Expert judgment was also used as a more holistic method of judging source material difficulty. The results of these analyses were compared with two groups of student interpreters' performance on consecutive interpreting of the source materials into Mandarin Chinese. The participants' assessment of speech difficulty after the interpreting task was also compared with the other measures and the expert judgment. The quantifiable measures all failed statistically in predicting source material difficulty, possibly due to the very small sample size of the materials or to the fact that the materials were very similar in the aspects assessed by these measures. A trend emerged to suggest that information density and sentence length may be potentially useful indicators for predicting source material difficulty. It was also shown that source material difficulty affected the performance of lower-skilled interpreters more than that of higher-skilled interpreters.

Introduction

Practitioners, trainers and test-developers in translation and interpreting often have to deal with source materials with different levels of difficulty. Practitioners develop strategies to tackle difficult elements in a source material. Trainers and testers gauge source material difficulty to match training and testing objectives. In these cases, judgment of the source material is often guided by intuition and experience rather than by systematic explorations. In the training and testing of

translation and interpreting, it is common practice that the selection of source materials relies solely on the judgment of an individual trainer or tester. In some testing situations such as the professional examinations of a translation and interpretation training institute in which the first author has participated, when a group of jurors judge and discuss source material difficulty or suitability, there is often a lack of consensus. This situation is not unlike what has been observed and studied in language testing. When a group of experts are called upon to judge text difficulty, it is usually difficult to reach a consensus as each individual may focus on different elements of the text (Alderson 1993; Fulcher 1997).

This lack of consensus can be particularly pronounced in interpreting because of the fleeting nature of the task. A text that reads smoothly in print may be difficult to comprehend when presented orally. The working conditions of an interpreter often affect how difficult a speech is perceived to be and how well it is interpreted. In addition, the mode of interpreting, be it consecutive or simultaneous, may also make interpreting a particular speech more or less difficult. Added to the complexity is the necessary consideration of the interpreter's background knowledge and the extent to which she prepares for an interpreting task.

The interaction of these factors not only makes judging interpreting source materials difficult in practice, but also makes theorizing about input difficulty a complex task. Campbell (1999) sees the question of source text difficulty as particularly complex and notes that a lack of suitable models in translation studies may have prevented translation educators from effectively incorporating the notion of difficulty into courses and tests.

Indeed, difficulty has also been examined in a rather patchwork manner in interpreting studies. Some "input variables" (Pöchhacker 2004: 126) that make an interpreting task difficult have been identified and studied. Among these factors, those related to speaker characteristics and working conditions such as input speed, intonation and background noise are observed to make the interpreting task more difficult (e.g., Gerver 1969/2002, 1974; Lee 1999a; Pio 2003; Tommola & Lindholm 1995). Some source material characteristics are also shown to have negative effects on interpreting performance. These include information density (e.g., Barik 1973, 1975; Dillinger 1994; Lee 1999a, 1999b), syntactic complexity and lexical difficulty of the source material (e.g., Darò et al. 1996; Tommola & Helevä 1998).

While factors such as speed, intonation and noise can easily be monitored and controlled in test development, the intrinsic elements of the source material are difficult to control due to a lack of quantifiable measures. This situation is problematic in testing for the purpose of certifying interpreters, as a lack of consistent control of test difficulty can lead to a slide in standards and eventually jeopardize the credibility of the certification.

One may argue against the need for establishing quantifiable measures for judging interpreting source material difficulty due to the interaction of different factors. However, it is for the same reason that, borrowing Campbell's words (1999: 34), "the problem cannot be approached holistically," and it is worth contemplating the possibility of the source text being an "independent source of translation difficulty." This pursuit not only has theoretical merit, but also has profound practical value in training and, in particular, testing.

The exploratory study reported here was among the pilot studies of a research project that the first author and her team undertook for Taiwan's National Institute for Compilation and Translation, with the aim of establishing a certification program for translators and interpreters in Taiwan.[1] The study focused on finding quantifiable measures for estimating and predicting the difficulty of English source materials for consecutive interpreting. The measures chosen for analysis included the readability level based on a readability formula, information density and new concept density based on propositional analysis.

Readability was chosen as a potential indicator because word length and sentence length, two elements commonly factored in calculating the readability level, can be used to gauge lexical difficulty and syntactic complexity. Information density was determined by the use of propositional analysis because of this method's preciseness in representing the meaning units of a text. New concept density, also based on propositional analysis, was used because a higher redundancy level (i.e., less new information) in the source material has been suggested as a factor that makes an interpreter's task easier (Déjean Le Féal 1982).

In addition to the above-mentioned quantifiable measures, the pooled judgment of a group of ten experts was also used as a more holistic method of judging source material difficulty. The results of these analyses were then compared with the scores of student interpreters' consecutive interpreting of the English source materials into Mandarin Chinese. The interpreters' own assessment of input difficulty after the interpreting task was also analyzed, particularly in relation to expert judgment as a holistic method of assessing source material difficulty.

Method

Participants

Two groups of students from a Taiwanese university participated in the study. The first group was composed of four graduate students of interpreting (three females and one male), who, by the time of the experiment, had received training in consecutive interpreting for about half a year. The other group was composed of seven

undergraduate English-major seniors (four females and three males), who, except for one without any training in interpreting, had taken a weekly two-hour consecutive interpreting class for about three months. All participants were in their early or mid twenties and had Mandarin Chinese as their first language and English as their second language.

The participants in this study are representative of the type of people who take the new Taiwanese interpreting certification examinations which currently only test consecutive interpreting. These participants are also typical of a rather large population of interpreting learners in Taiwan as most English departments at colleges and universities offer training in consecutive interpreting.

Materials

Three non-technical English texts were chosen to be the experimental materials by two experienced interpreting practitioner/trainers with Mandarin Chinese A and English B. The texts were considered to be prototypical talks Taiwanese interpreters encounter in assignments where the consecutive mode of interpreting is used.

The first text is a talk given at *Computex*, an international computer exhibition held in Taipei (*Computex*). The speaker talks about the theme and activities of the exhibition. The second text is an educational talk about user agreements that precede the installation of free software programs from the Internet (*Eula*). In the third text, the speaker talks about how different parties view a model of partnership between government and private companies called "public-private partnership" (*PPP*) (see Appendix for experimental materials).

All the texts were read out by a native English speaker and digitally recorded. The speed of the recorded speeches was adjusted to about 100 to 110 words per minute. Each speech was further divided into segments of 45–65 words, or in terms of speech time, 25–40 seconds (see Table 1). As the experiment was not meant for evaluating the participants' note-taking skills in consecutive interpreting, the division of the speeches into smaller segments allowed those participants who had not been sufficiently trained in consecutive interpreting note-taking skills to rely on their memory or their own method of note-taking to recall the content of the speeches.

Preliminary ranking of source material difficulty

The two experts who chose the experimental materials judged the difficulty of the three texts independently. They later compared their assessments and together ranked the three source materials in terms of their difficulty levels. *Computex* was judged to be the easiest and *PPP* the most difficult. The difficulty levels were assessed and determined based on the following eight categories that emerged from

Table 1. Source materials and interpreting segments

Title	Segments	No of words	Duration (sec.)
Computex (A)	whole text	178	104
	A1	57	33
	A2	62	38
	A3	59	33
Eula (B)	whole text	230	122
	B1	64	36
	B2	51	27
	B3	65	33
	B4	50	26
PPP (C)	whole text	204	112
	C1	48	27
	C2	44	25
	C3	51	30
	C4	61	30

the discussion of the two experts: words, syntactic structure, information density, coherence, logic, clarity, abstractness, and required background knowledge.

Methods of judging source material difficulty

As mentioned before, four methods were used to estimate and predict the difficulty of the three source materials. They are the readability levels based on the Flesch Reading Ease formula, information density based on propositional analysis, new concept density based on propositional analysis, and expert judgment.

Flesch Reading Ease formula for readability

Readability formulas are created out of an effort to find statistical correlations between "objectively observable features" of texts and the reading levels of readers (Davison & Green 1988: 1). The text features incorporated in readability formulas usually include average sentence length, and word difficulty, either based on the average number of syllables or occurrence of high-frequency words (Davison & Green 1988: 2). The underlying assumptions of using sentence length as an element in the readability formula is that the longer a sentence is, the more clauses it contains and thus the more complicated the sentence structure is (Anderson & Davison 1988; Kintsch & Miller 1984). As a longer sentence may generally contain

more information, the average sentence length can also be used as an indicator for information density in a text (Dam 2001: 30–31).

The Flesch Reading Ease formula, a reading difficulty measure designed for adults (Harrison 1980, cited in Fulcher 1997: 499) and one of the most frequently used readability formulas, was chosen for this study. The formula is based on word length (number of syllables per word) and sentence length (number of words per sentence). The scores range from 0 to 100, with more difficult texts having lower scores. The Flesch formula has been shown to correlate at around 0.64–0.70 with such measures as cloze and teacher judgment (Harrison 1980, cited in Fulcher 1997: 501). In a study where the difficulty level of certain sentences was adjusted according to the Flesch-Kincaid Readability Tests,[2] participants' performance in simultaneous interpreting was negatively affected by the more difficult sentences (Liu et al. 2004).

Propositional analysis for information density and new concept density

Propositional analysis is used to represent the content of a text in meaning units, expressed in the form of a list of propositions. A proposition is the smallest unit that is meaningful (Solso 1998: 259–260). A typical proposition is composed of a predicate and one or more arguments, with the predicate serving to specify the relationship among the arguments (Kintsch & van Dijk 1978). Studies have shown that texts containing the same number of words but more propositions take longer to read (Kintsch & Keenan 1973; Kintsch et al. 1975). Furthermore, it was also shown that with the number of words and propositions controlled, a text containing more new concepts (i.e., new arguments) took more time to read and was more poorly recalled (Kintsch et al. 1975).

In this study, propositional analysis was performed according to the guidelines specified in Bovair & Kieras (1985). The list of propositions for each source material served two purposes: first to calculate the density of information of each source material, and second, to serve as scoring units for the rating of interpreting performance. In its first application, the proportion of the number of propositions to the number of total words was calculated for each source material to determine its information density: the higher the score, the denser the information. In addition, the proportion of the number of new arguments to the total number of propositions in each source material was calculated to indicate the density of new concepts: the higher the score, the denser the new concepts.

Expert judgment

Expert judgment is widely used in language testing as well as in translator and interpreter testing. It is often assumed that experts are able to predict test difficulty in advance of a test administration, particularly in translator and interpreter

testing where test piloting is difficult to do. We would expect that experts, with their long experience in teaching and testing, would have internalized a notion of difficulty in relation to how candidates are expected to perform. However, expert judgment has often been shown to be very unreliable. In some studies, the lack of agreement not only exists among experts (test writers and examination markers), and between item statistics and expert judgment, but also shows in intra-rater unreliability (Alderson 1993; Fulcher 1997). This situation makes the approach of using expert judgment in content validation highly questionable. The absence of commonly shared criteria and different weightings given to each criterion may contribute to this problem. Therefore, having experts work in a team can be beneficial for the purpose of reaching an agreement (Fulcher 1997: 503).

In this study, a group of ten experts composed of professional interpreters, interpreter trainers, as well as reading and language testing experts was invited to fill out a questionnaire that asked about different aspects of the testing of consecutive interpreting. As part of the questionnaire, items about source material difficulty were also included. These experts were asked to make an overall judgment of text difficulty on a five-level Likert scale, with 1 being "very easy" and 5, "very difficult." In an effort to allow more guided judgment, questions incorporating the eight categories used by the two experts who chose and ranked the source materials were also included in the questionnaire. The eight categories, as described earlier, include some criteria that are similar to the ones assessed by the quantifiable measures used in this study, such as information density, word difficulty and syntactic difficulty. Other criteria not assessed by the quantifiable measures are also included and, as a group, represent the more holistic nature of expert judgment. The experts were also asked to assess each of these eight categories by marking their choices on a five-level Likert scale. Assessment made during these two stages allowed us to compare the results of expert judgment without a guideline with those based on specific criteria.

Procedure

All eleven participants performed the interpreting task individually. The order of interpreting the three source materials was determined by a blocking strategy for each group to distribute practice and fatigue effects. Each participant sat in a simultaneous interpretation booth and was first asked to read a short statement that included a one-sentence description of the topic and content of the source material to be interpreted, the length and number of segments of that specific source material, and a reminder that they could take notes using the paper and pens provided. The participants were also allowed time to ask questions about the

experimental procedure. Next, they listened to the first recorded source material through headphones and their interpretation renditions were recorded onto cassette tapes and later converted to digital sound files. Interpreting sessions of the second and third source materials followed the same procedure. The participants were allowed short breaks between sessions if they so desired. No time limits were imposed on the participants so that they could proceed at their own pace in interpreting the three speeches. All participants were observed to take notes while listening to the three source materials.

At the end of the last session, each participant was asked on a voluntary basis to fill out a questionnaire about input difficulty, the quality of their own performances, input speed, the length of each interpreting segment, and the necessity of note-taking. For each question, the participants had to mark their choice on a five-level Likert scale. The whole process lasted between 30 and 60 minutes for each participant.

Rating of interpreting performance

The quality of the interpreting performance was evaluated by using a proposition-based rating method. It was done by calculating the percentage of propositions of each source material correctly interpreted. Only fidelity of the interpretation renditions was considered in the rating as accuracy is a more clear-cut criterion for determining the effect of input difficulty on the interpreting performance.

All 33 interpretation recordings were transcribed verbatim. For each source material, the interpretation transcripts were randomized so that the order of rating for each participant was different. The accuracy of all interpretation renditions was assessed by checking how closely each transcribed interpretation rendition matched the propositions of each source material. A score of 1 was given when the meaning of a proposition was correctly interpreted, and a score of 0 when it was wrong. Two native Chinese speakers with graduate-level interpreter training served as raters.

First, the two raters did a trial rating session individually on the interpretation renditions of three randomly chosen participants. After discussing the results of their trial rating, they agreed on the following principles before proceeding:

1. Credit is given to renderings that may not be equivalents but nonetheless represent the meaning of the original propositions;
2. Renderings that do not follow the original order of propositions in a sentence are not penalized, due to word order difference between English and Chinese;
3. Added information that is not in the original list of propositions is disregarded and not penalized;

4. Erroneous renderings of the same proposition are penalized only once when they first appear.

To assure better consistency in rating, all participants' renditions of the same segment of a particular speech were rated before proceeding to the next segment. The final score for each interpreting performance was the average of the scores given by the two raters, calculated by dividing the number of correctly interpreted propositions by the total number of propositions in each source material.

Statistical Analysis

Inter-rater reliability was examined using the Kappa statistic. A two-way analysis of variance (ANOVA) was used for analyzing the effects of the three source materials and the two groups of participants. Pearson's correlation coefficients were calculated to determine the correlations of readability score, information density, new concept density and expert judgment with the scores of the interpreting performance.

Results and Discussion

Estimated difficulty of source materials

The difficulty levels of the three source materials, as assessed by the four indicators, are presented in the following three tables. As can be seen in Table 2, the readability levels represented by the Flesch Reading Ease scores show *Computex* to be the most difficult, and *PPP*, the easiest.

Table 2. Readability of source materials in Flesch Reading Ease scores

	Word length	**Sentence length**	**Reading Ease score**	**Difficulty level**
Computex	5	19.7	35.6	Difficult
Eula	4.4	19.2	55.3	Fairly difficult
PPP	4.4	22.5	60.1	Standard

Note. A Flesch Ease score of 0–30 is considered "very difficult," 30–50 "difficult," 50–60 "fairly difficult," 60–70 "standard," 70–80 "fairly easy," 80–90 "easy," and 90–100 "very easy" (Flesch 1948: 230).

Table 3 shows that the three source materials are quite comparable in information density, with *PPP* judged to contain the densest information. The result of the new concept density is very different. *Computex* has far more new arguments than the other two source materials, and thus should be the most difficult one to interpret.

Table 3. Information density and new concept density of source materials

	Number of words	Number of propositions	Information density (%)	Number of new arguments	New concept density (%)
Computex	178	76	**42.70**	63	**82.89**
Eula	230	96	**41.74**	37	**38.54**
PPP	204	88	**43.14**	52	**59.09**

Expert judgment exhibits a different result from that of the readability formula and propositional analysis. The score of overall difficulty and the average score of the eight criteria both show *Computex* to be the easiest, and *PPP*, the most difficult. In fact, *PPP* was rated as the most difficult in every category with the exception of coherence and logic.

Table 4. Expert judgment of difficulty of source materials

	Computex	*Eula*	*PPP*	Notes
Overall difficulty	**2.00**	**3.10**	**3.70**	5: most difficult
1. Word difficulty	2.00	2.60	3.00	5: most difficult
2. Syntactic difficulty	1.90	2.80	3.50	5: most difficult
3. Information density	2.90	3.30	3.50	5: most dense
4. Coherence	2.00	2.80	2.40	5: least coherent
5. Logic	1.80	2.60	2.50	5: least logical
6. Clarity	1.70	2.40	2.50	5: least clear
7. Abstractness	1.40	2.60	3.20	5: most abstract
8. Knowledge difficulty	2.10	3.10	3.70	5: most difficult
Average of 1 to 8	**1.98**	**2.77**	**3.03**	5: most difficult

Cronbach's alpha was calculated to measure the internal consistency reliability of expert judgment. The results showed that the experts were very consistent at judging *Computex* and *Eula* with the Cronbach's α at 0.87 and 0.82 respectively. However, internal consistency was much lower for the judgment of *PPP*, $\alpha = 0.45$, indicating much greater disagreement among experts in judging this text.

Rating results of interpreting performance

Inter-rater reliability was first examined. Pearson's chi-square test showed correlation in the two raters' ratings, $\chi^2 = 2135.16$ ($p = .000$). A strong inter-rater agreement of $K = 0.86$ was further obtained using the Kappa statistic.

Interpreting performance of *Eula* received the highest score at 49.53%. A comparable but slightly lower score of 47.19% was given to *Computex*, and *PPP*

received the lowest score at 39.98%. Judging by the participants' performance as a group, *Eula* seems to be the easiest material and *PPP*, the most difficult.

Graduate students outperformed undergraduates on all three source materials (see Table 5).

Table 5. Rating results of consecutive interpreting performance of two groups of participants

	Graduates (*N* = 4)		**Undergraduates (*N* = 7)**	
	Mean	***SD***	**Mean**	***SD***
Computex	67.93	16.27	35.34	12.97
Eula	59.64	08.70	43.75	08.67
PPP	62.36	15.40	27.19	11.07

A two-way mixed-model ANOVA showed a significant main effect of "group" at $F(1, 9) = 20.26$, $p = .001$, but the main effect of "material" was not significant, $F(2, 18) = 2.24$, $p > .05$. A significant interaction effect of "material" and "group" was observed at $F(2, 18) = 3.89$, $p < .05$. One-way ANOVA further showed that graduate students performed significantly better than undergraduates on all three source materials, $F(1, 9) = 13.49$, $p < .01$ for *Computex*, $F(1, 9) = 8.53$, $p < .05$ for *Eula*, and $F(1, 9) = 19.57$, $p < .01$ for *PPP*. Further analysis on repeated measures showed that graduate students did not perform differently in interpreting the three source materials, $F(2, 6) = 1.12$, $p > .05$, indicating that difficulty of these materials did not affect the graduate students' performance in consecutive interpreting. The interpreting performance of the undergraduates, on the other hand, was affected by the different source materials they interpreted, $F(2, 12) = 6.35$, $p < .05$. Further analysis, using the Scheffe test, showed that their performance on *Eula* was significantly better than that on *PPP*, $p < .05$, while the other pair-wise comparisons did not reach significance.

Participant judgment

Nine of the 11 participants filled out the participant questionnaire. As can be seen in Table 6, their assessment of speech difficulty shows that *PPP* was considered the most difficult (3.78), while *Eula* and *Computex* are similar in their difficulty levels (2.78 and 2.56 respectively). Participants' evaluation of their own performance, though generally rated low, shows *Eula* as being the best (2.56), followed by *Computex* (2.44), and *PPP* being the worst (1.78). Participants did not seem to attribute task difficulty to the speed of the source materials or the length of each interpreting segment, both rated at 3. However, note-taking seemed to be considered quite necessary (3.75), despite the limit of 40 seconds set for the length of each segment.

Table 6. Results of participant questionnaire

	Computex	*Eula*	*PPP*	**Notes**
Input difficulty	2.56	2.78	3.78	5: most difficult
Self evaluation of performance	2.44	2.56	1.78	5: best
Input speed		3.00		5: very fast
Segment length		3.00		5: very long
Necessity of notes		3.75		5: very necessary

Note. Numbers for input speed, segment length and necessity of notes represent scores for all three materials.

Correlation of estimated difficulty levels and interpreting performance

We compared the difficulty levels assessed by different quantifiable measures and expert judgment with the scores of interpreting performance to examine how effective each indicator was in predicting source material difficulty. Pearson's correlation coefficients were calculated for each indicator and the scores of interpreting performance (see Table 7). A significant negative correlation coefficient should indicate that an indicator was effective in judging source material difficulty, that is, the higher the difficulty level the lower the interpreting score.

Table 7. Correlation coefficients of consecutive interpreting performance and difficulty levels as assessed by different indicators

		r	*p*
Readability formula	Flesch Reading Ease score	.46	.70
Propositional analysis	Information density	−.88	.36
	New concept density	−.28	.82
Expert judgment	Overall	−.60	.59
	Word difficulty	−.64	.53
	Syntactic difficulty	−.67	.53
	Information density	−.58	.61
	Coherence	.24	.85
	Logic	−.17	.89
	Clarity	−.39	.75
	Abstractness	−.58	.61
	Knowledge difficulty	−.62	.58
	Average of 8	−.50	.67

Possibly due to the extremely small sample size of the three source materials or the fact that the materials are very similar in the aspects assessed by the different

indicators, none of the correlation coefficients of difficulty and performance turned out to be significant. In the case of the readability formula, the correlation coefficient came out positive, $r = 0.46$, indicating a peculiar situation where the more difficult a source material is, the better the performance in interpreting. Among all other indicators, overall information density assessed by propositional analysis has the highest negative correlation coefficient at $r = -0.88$, despite a lack of significance. However, new concept density and scores of interpreting performance only show an insignificantly low correlation at $r = -0.28$. Most of the expert judgment indicators show an insignificantly low to moderate correlation levels.

Despite a lack of significance, we tried to examine the trend of the use of different indicators in correctly predicting source material difficulty by ranking the different difficulty levels (see Table 8).

Table 8. A Ranking comparison of different difficulty levels and consecutive interpreting performance

	Readability	Information density	New concept density	Expert judgment	Participant judgment	Interpreting performance
Computex	3	2	3	1	1	2
Eula	2	1	1	2	2	1
PPP	1	3	2	3	3	3

Note. Difficulty level: 3 — most difficult; interpreting performance: 3 — worst.

Information density and new concept density

As can be seen in Table 8, the only indicator that shares the same ranking as the interpreting performance is information density. *PPP*, of which the interpretation renditions were consistently rated the poorest, was judged to be the most difficult by three indicators — information density, expert judgment, and participant judgment. This result seems to suggest that information density may be one of the most important factors that can affect interpreting performance. It is also possible that information density is the most perceivable difficulty factor.

Despite presenting a different result from that of information density, new concept density also shows that *Eula* is the easiest source material and thus corresponds to *Eula*'s best performance. As mentioned before, more new concepts result in longer reading time and poorer recall (Kintsch et al. 1975). In our study, *Computex* contains far more new concepts than *Eula* (with density levels at 82.89% vs. 38.54% respectively as shown in Table 3) but did not result in a much poorer performance in consecutive interpreting. This outcome can be interpreted in the following ways: Firstly, as a relatively easy text (judged by experts as the easiest, at 2.00 in overall difficulty as shown in Table 4), *Computex*'s abundance of new

concepts did not seem to constitute a strong enough factor to affect the quality of interpreting. For example, *Computex*'s new concepts appear at one point as a rather long list of closely related computer terms (see A2 in Appendix), which, due to their use in everyday life, did not seem to cause much difficulty in the participants' interpreting performance. Secondly, recall, as a way to retrieve stored information, can be quite different from consecutive interpreting, which is often aided by notes taken during the process. All participants in our study chose to take notes and most of them considered note-taking necessary (3.75 out of a highest score of 5 as shown in Table 6) to produce a good performance. If the participants were asked to do recall instead of consecutive interpreting, it is quite possible that more information would be lost. Thirdly, by the time the interpreter starts interpreting, the new concepts may have already been integrated in the overall structure of the knowledge base built for the source material and thus do not cause much difficulty with their newness. It is possible that the result may be quite different with simultaneous interpreting, as interpreters need to deal with new information as it arises.

Expert judgment and participant judgment

As can be seen in Table 8, the rankings of the indicators other than information density do not correspond to the ranking of the interpreting performance. Also, they do not correspond to each other in most cases. The only exception is between expert judgment and participant judgment, which share the same ranking. However, a closer look at the results of expert judgment and participant judgment reveals that there is a rather big gap between the difficulty scores of *Computex* and *Eula* given by the experts (2.00 vs. 3.10 in overall judgment and 1.98 vs. 2.77 in the eight-category average score, shown in Table 4), while the difference between the two scores given by the participants is almost negligible (2.56 vs. 2.78, shown in Table 6). In further comparing these results to the interpreting performance scores, we see that the participants' judgment of difficulty seems to better reflect the outcome of their interpreting performance, despite the difference in ranking. The differences between the interpreting scores of *Computex* and *Eula* are minor in the proposition rating (47.19% vs. 49.53%), as well as in the participants' own evaluation of their performance (2.44 vs. 2.56, shown in Table 6).

It is interesting to note that experts made their judgment by reading the original texts of the source materials, while the participants judged input difficulty immediately after interpreting the source materials. The results mentioned above can be explained by saying that certain features of a source material may be emphasized when the assessment is based on the written text or when it is done without actually interpreting the material. When examining how each of the eight categories was rated by the experts, we see that *Computex* and *Eula* exhibit the most pronounced differences in the categories of "abstractness" and "knowledge

difficulty," with a difference of over 1 point (out of 5) in both cases (see Table 4). It is quite possible that these two features are more elusive in nature and that judges may have a tendency to define them differently.

Aside from the issue of source material difficulty, it is interesting to note that the participants' evaluations of their own interpreting performance are quite accurate in the sense that they match the rating results in ranking (see Tables 6 and 8). Considering that the participants in this study were all students, this result makes pedagogical sense, in that learners of interpreting may be more actively engaged in the learning and evaluation process.

Readability

Among all the indicators, readability is the only one that does not involve information in its calculation. This may explain why the ranking of readability, using the readability formula, is the most dissimilar to that of the interpreting performance. *Computex* was judged to be the most difficult, while *PPP* was seen as easiest. Among the three source materials, *Computex* has the longest words (see Table 2), possibly due to the inclusion of a list of multisyllabic, computer-related terms in the second paragraph. However, these terms are very common in an educated adult's lexicon and, judging by the participants' interpreting performance, did not seem to cause much difficulty.

The conception and use of readability formulas have received much criticism for their limitations in reflecting some important factors pertaining to the text and the reader (Davison & Green 1988; Fulcher 1997). For example, important factors such as conceptual complexity, text organization, or reader's knowledge and strategies are ignored in the readability formulas (Fulcher 1997: 501). Despite the fact that the calculation based on a readability formula only reflects the difficulty of the surface elements of a text and not its "conceptual complexity" (Kintsch & Miller 1984: 221), the inclusion of word length and sentence length has its cognitive underpinnings in how information is processed. Word length can affect the speed of recognition of a word (Smith & Taffler 1992: 86) and may take longer to rehearse subvocally (Baddeley & Logie 1999). Chincotta and Underwood (1998) have demonstrated that in the case of professional simultaneous interpreters, the word-length effect is not completely eliminated even though subvocal rehearsal is partly or completely prevented. Therefore, the recall of shorter words in the target language is still expected to yield better results. Sentence length has also been suggested to cause an impact on "immediate processing capacity" and to affect the recall of words (Anderson & Davison 1988: 42; Smith & Taffler 1992: 86). These two factors seem to be particularly relevant to the process of interpreting, which, more than reading, seems to rely heavily on how fast information is processed and how working memory can be efficiently used.

It is interesting to note that *PPP*, with the longest average sentence length among the three source materials, is judged to be the easiest by the readability formula (see Table 2). In the Flesch Reading Ease formula, word length is apparently weighted more heavily than sentence length.[3] It is possible that in interpreting, sentence length, which often reflects sentence complexity and possibly denser information, may be a more relevant factor than word length in affecting how information is processed. In this sense, the element of sentence length in the formula, instead of the complete formula, may be a more reliable indicator of source material difficulty. However, it should be noted that not all long sentences are complex. A long sentence containing several coordinate clauses may not be as complex as a shorter sentence with an embedded relative clause. Likewise, a long sentence with a subordinating conjunction "because" may be easier to comprehend than two short sentences with a covert cause and effect relationship (Anderson & Davison 1988). As this distinction is usually not addressed in a readability formula, judgment of difficulty based on sentence length should not be blindly accepted without further analyzing the complexity of a sentence.

Another factor that may affect the usefulness of the readability formula in judging source material difficulty for the task of interpreting is its suitability for judging spoken materials. Past studies have not pointed to a consistent correlation between listenability and readability (Klare 1963, cited in Dubay 2004: 46). However, one of the very few studies in interpreting studies that used a readability formula in distinguishing experimental material difficulty has shown that sentence difficulty so judged had a significant effect on participants' performance in English to Mandarin Chinese simultaneous interpreting (Liu et al. 2004). However, it is quite possible that different modes of interpreting may produce different results because of the way information is processed. In simultaneous interpreting, the more linear information processing can result in more local comprehension of the input, while in consecutive interpreting, more global comprehension of the input is required. In this sense, it is possible that a source material judged to be more difficult by a readability formula (more difficult surface structure) can affect the performance in simultaneous interpreting more than in consecutive interpreting.

Like the other factors examined in this study, readability needs to be studied in a more controlled way using a greater variety of texts. Meanwhile, the use of readability formulas in interpreting, as in reading, should not be seen as a shortcut to solve the problems in selecting materials for pedagogic or testing purposes (Fulcher 1997: 510). In addition, the interpretation of the calculation results of readability formulas should be done with great caution.

General discussion and conclusion

The current study was motivated by the need for better control of examination standards of a certification exam for interpreters, since a more objective method of judging and controlling source material difficulty will greatly benefit this endeavor.

We used a combination of quantifiable (and thus more objective) measures that examined different aspects of the source material, such as word length, sentence length, information density and new concept density. We compared the results of these measures with the results of expert judgment. The difficulty levels assessed by these indicators were later compared with participants' performance in consecutive interpreting and their assessment of input difficulty.

Statistically, the indicators we studied have failed individually in predicting source material difficulty. The very small sample size of three source materials might have caused the lack of significance. It is also possible that the three source materials chosen for this study were all similar in the aspects examined by the indicators and thus could not be distinguished from each other in terms of their difficulty levels. The lack of statistical significance, however, does not necessarily suggest that the factors represented by these indicators are unrelated to source material difficulty. Indeed, many aspects in the results of this study show interesting trends upon which we can form hypotheses for further investigation.

We found that the difficulty levels determined by a readability formula based on word length and sentence length are not reliable. Word length may correspond generally to word complexity and abstractness (Flesch 1948: 226), but may not constitute a difficulty factor in consecutive interpreting where words and ideas are jotted down as notes to be expressed in a different language at a later time. When immediate processing of words is required, as in the task of simultaneous interpreting, it is possible that longer words may prove more difficult to process.

Sentence length, the other element in the readability formula chosen for this study, may be a relevant difficulty factor for the task of interpreting as the source material with the longest sentences in this study did elicit the poorest interpreting performance. As this particular material (*PPP*) is also considered to contain the densest information according to the ranking (see Table 8), further research employing a variety of texts is needed to determine if longer sentences with greater complexity, or denser information, or both make a source material difficult. It is also possible, as mentioned earlier, that longer sentences generally contain more information (Dam 2001). Naturally, this conclusion is only valid when redundancy of information is taken into consideration. However, our choice of measurement for the lack of redundancy (old information), i.e. new concept density, failed to correspond neatly with this consideration. While the source material with the

lowest new concept density (*Eula*) turned out to receive the best score in interpreting performance and thus seemed to be the easiest to interpret, the one with the densest new concepts (*Computex*) did not seem to be the most difficult to interpret as its interpretation renditions did not receive the lowest score.

Although it is used for pedagogical and testing purposes in language education as well as in translation and interpreting, expert judgment has been shown to be unreliable (Alderson 1993; Fulcher 1997). In the present study, the pooled judgment of ten experts was used to judge source material difficulty. The result of expert judgment seems to correlate at no more than medium level with the scores of interpreting performance.

The literature has shown that the great discrepancy in expert judgment comes from the different aspects each individual emphasizes in a text (Fulcher 1997). We asked the experts in this study to also judge the source materials based on eight categories so as to allow a more guided judging experience. The overall judgment and category-guided judgment showed the same result in ranking the source materials but differed in the gaps between the scores given to the three materials. The differences in scores in the guided judgment are smaller than those in the overall judgment. It is possible that when judgment is done holistically and without guidelines, certain aspects in a source material may attract more attention and their effects may be skewed.

Individual differences such as previous experience, depth of background knowledge, and domain skills have constantly been shown to be an important factor in whether a text or a task is perceived to be easy or difficult (Anderson & Davison 1988). In our study, the graduate students who received lengthier and more intensive training interpreted the three source materials equally well, as can be seen in the lack of statistically significant differences in the scores of the three source materials. The performance by undergraduate students, on the other hand, showed statistically significant differences in two of the three source materials. Like reading, where the reader and the reading material interact to determine the success of the outcome (Anderson & Davison 1988; Bailin & Grafstein 2001), success in interpreting a speech depends not only on the material itself, but also on the interpreter. In this study, note-taking skills might be a particularly important factor in the student interpreters' performance in consecutive interpreting. While all of the participants in the study took notes, the relative efficiency and effectiveness of their note-taking could also be a confounding factor.

This study, being one of the pilot studies of a larger research project and exploratory in nature, yielded rather few results, due to its use of only three research materials and two very small and heterogeneous groups. Future studies incorporating a greater number and variety of materials or even materials specially designed for

more controlled observations are needed for clearer results. The participation of a larger number of more homogenous participants would also be beneficial. Nonetheless, we hope that the results of this study have helped highlight an important research direction that has not received enough attention in interpreting studies.

Notes

* This study was supported by a grant from Taiwan's National Institute for Compilation and Translation in 2006. The content of this article is based on part of the final report of the first author and her team's research project "Establishment of a national standard for the evaluation of translation and interpreting — 3rd phase" (Liu et al. 2006) and the second author's MA thesis on the subject (Chiu 2006). Some of the results of the study have been presented at the Newcastle University Conference on Interpreter and Translator Training and Assessment, Sept. 9–10, 2007, Newcastle, England.

1. Taiwan had its first certification examinations for translators and interpreters in December 2007. The "Chinese and English Translation and Interpretation Competency Examinations" are administered by the Taiwanese Ministry of Education. Currently, only consecutive interpreting is tested in the interpretation competency examinations.

2. The Flesch Reading Ease formula is part of the Flesch-Kincaid Readability Tests, which also include the Flesch-Kincaid Grade Level.

3. The Flesch Reading Ease formula reads as follows (Flesch 1948):

206.835 – 84.6 × ASW (average number of syllables per word) – 1.015 × ASL (average sentence length).

4. We would like to thank Shu-Pai Yeh, Chia-Ling Cheng and Hong-Yu Huang for their assistance in this study.

References

Alderson, J. C. (1993). Judgments in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research: Selected papers from the 1990 language testing research colloquium*. Sterling, VA: TESOL, 46–57.

Anderson, R. C. & Davison, A. (1988). Conceptual and empirical bases of readability formulas. In A. Davison & G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidere*d. Hillsdale, NJ: Erlbaum, 23–53.

Baddeley, A. D. & Logie, R. H. (1999). Working memory: The multiple-component model. In P. Shah & M. Akira (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge, UK: Cambridge University Press, 28–61.

Bailin, A. & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language and Communication* 21, 285–301.

Barik, H. C. (1973). Simultaneous interpretation: Temporal and quantitative data. *Language and Speech* 16, 237–270.

Barik, H. C. (1975). Simultaneous interpretation: Qualitative and linguistic data. *Language and Speech* 18, 272–297.

Bovair, S. & Kieras, D. E. (1985). A guide to propositional analysis for research on technical prose. In B. K. Britton & J. B. Black (Eds.), *Understanding expository text*. Hillsdale, NJ: Erlbaum, 317–362.

Campbell, S. (1999). A cognitive approach to source text difficulty in translation. *Target* 11 (1), 33–63.

Chincotta, D. & Underwood, G. (1998). Simultaneous interpreters and the effect of concurrent articulation on immediate memory. *Interpreting* 3 (1), 1–20.

Chiu, Y-H. (2006). *Assessing input difficulty in interpretation: An experiment of English to Chinese consecutive interpretation* (in Chinese). Unpublished MA thesis, Fu Jen University, Taiwan.

Dam, H. V. (2001). On the option between form-based and meaning-based interpreting: The effect of source text difficulty on lexical target text form in simultaneous interpreting. *The Interpreters' Newsletter* 11, 27–55.

Darò, V., Lambert, S. & Fabbro, F. (1996). Conscious monitoring of attention during simultaneous interpretation. *Interpreting* 1 (1), 101–124.

Davison, A. & Green, G. M. (1988). Introduction. In A. Davison & G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidere*d. Hillsdale, NJ: Erlbaum, 1–4.

Déjean Le Féal, K. (1982). Why impromptu speech is easy to understand. In N. E. Enkvist (Ed.), *Impromptu speech: A symposium*. Åbo: Åbo Akademi, 221–239.

Dillinger, M. (1994). Comprehension during interpreting: What do interpreters know that bilinguals don't? In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the gap: Empirical research in simultaneous interpretation*. Amsterdam: John Benjamins, 155–189.

Dubay, W. H. (2004). The principles of readability. http://www.impact-information.com (accessed 13 June 2005).

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology* 32, 221–233.

Fulcher, G. (1997). Text difficulty and accessibility: Reading formulae and expert judgment. *System* 25, 497–513.

Gerver, D. (1969/2002). The effects of source language presentation rate on the performance of simultaneous conference interpreters. In F. Pöchhacker & M. Shlesinger (Eds.), *The interpreting studies reader*. London: Routledge, 53–66.

Gerver, D. (1974). The effects of noise on the performance of simultaneous interpreters: Accuracy of performance. *Acta Psychologica* 38, 159–167.

Harrison, C. (1980). *Readability in the classroom*. Cambridge: Cambridge University Press.

Kintsch, W., & Keenan, J. M. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology, 5*, 257–274.

Kintsch, W., Kozminsky, E., Streby, W. J., McKoon, G. & Keenan, J. M. (1975). Comprehension and recall of text as a function of content variable. *Journal of Verbal Learning and Verbal Behavior* 14, 196–241.

Kintsch, W. & Miller, J. R. (1984). Readability: A view from cognitive psychology. In J. Flood (Ed.), *Understanding reading comprehension: Cognition, language and the structure of prose*. Newark, DE: International Reading Association, 220–232.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review* 85, 363–394.
Klare, G. R. (1963). *The measurement of readability*. Ames: Iowa State University Press.
Lee, T. (1999a). Speech proportion and accuracy in simultaneous interpretation from English into Korean. *Meta* 44, 260–267.
Lee, T. (1999b). Simultaneous listening and speaking in English into Korean simultaneous interpretation. *Meta* 44, 560–572.
Liu, M., Chang, V., Lin, S-H., Chen, B-J. & Chiu, Y-H. (2006). *Establishment of a national standard for the evaluation of translation and interpreting — 3rd phase* (in Chinese) (National Institute for Compilation and Translation research report). Taipei: National Institute for Compilation and Translation.
Liu, M., Schallert, D. L. & Carroll, P. J. (2004). Working memory and expertise in simultaneous interpreting. *Interpreting* 6 (1), 19–42.
Pio, S. (2003). The relation between ST delivery rate and quality in simultaneous interpretation. *The Interpreters' Newsletter* 12, 69–100.
Pöchhacker, F. (2004). *Introducing interpreting studies.* London: Routledge.
Smith, M. & Taffler, R. (1992). Readability and understandability: Different measures of the textual complexity of accounting narrative. *Accounting, Auditing & Accountability Journal* 5, 84–98.
Solso. R. L. (1998). *Cognitive psychology* (5th edn). Boston: Allyn and Bacon.
Tommola, J. & Helevä, M. (1998). Language direction and source text complexity: Effects on trainee performance in simultaneous interpreting. In L. Bowker, M. Cronin, D. Kenny & J. Pearson. (Eds.), *Unity in diversity: Current trends in translation studies*. Manchester, UK: St. Jerome, 177–186.
Tommola, J. & Lindholm, J. (1995). Experimental research on interpreting: Which dependent variable? In J. Tommola (Ed.), *Topics in interpreting research*. Turku: University of Turku, 121–133.

Appendix: Experimental materials

Computex (A)
A1
Welcome back! I just checked into a lovely hotel in downtown Taipei, Taiwan after having spent fifteen hours on an airplane on a direct flight to Taipei from Amsterdam. The reason I am in Taipei obviously is Computex. This yearly conference was actually postponed from its original June 2003 timeframe due to the SARS threat in Asia.

A2
For those of you that aren't familiar with Computex let me briefly, by using keywords, describe what Computex is all about. Computex is all about computers, components, communications, peripherals, storage devices, software, etc. Basically a great opportunity for many of the Asian, but mostly Taiwanese and Chinese, manufacturers to show off their new products and technologies to attendees from around the world.

A3
Tomorrow morning at nine the conference starts with an opening speech by representatives from the organization behind Computex, CETRA, the China External Trade Development Council. With over 1200 exhibitors and close to 2500 booths I'm sure there will be lots to tell and to show you this week, so stay tuned and check back soon for a daily update.

Eula (B)
B1
I'm sure we've all been faced with installing a piece of software and having to accept a user agreement in order to complete the installation. Actually these user agreements have become so commonplace you'll be hard pressed to find software that does not require you to accept a user agreement. As a result I hardly give it much thought anymore and just click 'accept'.

B2
So why did I not bother to read them? Partly because these agreements usually don't limit me in the way I use the software. They're mostly meant to safeguard the developer's intellectual property. I'm fine with that, as I have no intention to use it other than how the developer intended.

B3
Because I mostly download my software instead of buying it in the store, it usually comes with a trial period, giving me time to evaluate it before I accept the user agreement. When I buy software in a store, things are different though. I obviously can't try the software out before I buy it, nor will the store give me a refund for software returns.

B4
So in fact, you have already accepted the user agreement when you decide to buy it. That's actually a violation of my rights as a consumer, I'd like to know what I agree to prior to buying the software and installing it, but most store policies do not allow it.

PPP (C)
C1
Today we will hear from colleagues who will not only share their experiences with us, but help us open our minds on ways to make PPP and BOT in Taiwan more efficient and effective than ever before. Perhaps I can start the ball rolling with three simple observations.

C2
You may look at PPP or BOT as a financing vehicle, a way to get the best from the private sector supporting the public, but other key stakeholders have very different perceptions — perceptions which could either slow down or even stop a project.

C3
For example, workers in the public sector may be concerned that bringing in the private sector in this way will automatically lead to job losses. They will not necessarily see that it is often only by PPP that you can develop new job-creating infrastructure which you otherwise could not have afforded.

C4
They may also think there is a loss of face. 'Are we not good at our jobs? Why do you need to bring in the private sector who just wants to get rich and go away?' These are issues you need to address for all PPP and BOT projects, not project by project as too many local and national governments do.