FURTHER EXPERIMENTS IN LANGUAGE TRANSLATION : A SECOND EVALUATION OF THE READABILITY OF COMPUTER TRANSLATIONS

H. Wallace SINAIKO
Institute for Defense Analyses
Science and Technology Division
Arlington, Virginia *)

George R. KLARE
Ohio University

ABSTRACT

Language translation by computer has been proposed as a solution to the backlog of training and operational manuals awaiting translation by more conventional means. This study reports one of a series of experiments to assess the quality of translations produced by human translators and computers. The type of material under study was technical text (i.e., maintenance manuals) and the translation was from English to Vietnamese.

Utility or readability of the translations was assessed by reading comprehension tests, the cloze procedure (in which readers filled in blanks where words had been systematically deleted) and a rating scale for judging clarity. Time to perform each of these tasks was also measured. The subjects were 141 Vietnamese Navy officer candidates and a control group of 57 U.S. Navy officer candidates. A 500-word passage, from a U.S. Navy casualty control instruction, was translated by computer into a rough (un-edited) and a finished (post-edited) version; also, highly competent human translators prepared a Vietnamese text. Some Vietnamese subjects served as controls and took all tests based on the English, or untranslated, version.

Major conclusions were: (1) Translations produced by highly qualified humans were consistently more comprehensible than those produced by computer, whether edited or un-edited; post-edited versions of computer produced text were more comprehensible than unedited ones; most

*) Now at the Smithsonian Institution, Washington D.C. 20560.

differences were not statistically significant; (2) Vietnamese Navy officer candidates were able to read text in English as well as its best Vietnamese version and their test scores were about as high as those of American control subjects. Reading speed was not affected by mode of translation.

I. Background

The problem

This paper reports a continuation of a series of experiments which have sought to measure the utility of technical English that has been translated into Vietnamese (Sinaiko and Brislin, 1970; Klare, Sinaiko, and Stolurow, 1971; Sinaiko and Klare, 1972). The present experiment is primarily a comparison of the readability of an English text with translations of it into Vietnamese by computer and high-quality human translators. By readability we mean comprehensibility of a resulting translation to a representative reader. Our data were derived from readability tests administered to Naval Officer Candidates—both U.S. Navy and Vietnamese Navy—at Newport, Rhode Island.

This experiment is similar to an immediately preceding study (Sinaiko and Klare, 1972), in the following regards:

1. Machine translations were done with the LOGOS I computer-based system.
2. Human translations, used as a control condition, were done by the same expert translators.
3. The overall measurement scheme included reading comprehension tests, intelligibility rating scales, and cloze procedure.

The experiments differed, however, in certain important ways as follows:

1. The subject matter of the present translation was based on a U.S. Navy casualty control instruction intended for use by operational units, while the earlier study used samples of text from an Air Force instructional manual, *Instrument Flying*.
2. Because the LOGOS I system has been in a continued state of development, certain procedural and software changes have been introduced (but not specified to us) between the first machine translations and the current ones.

3. One of our measuring techniques, the use of reading comprehension tests, has been modified from a fill-in format to multiple-choice test items.
4. Vietnamese reader-subjects differed between the experiments in that the first group, Air Force, was widely varying in flying training and exposure to English, while the second group, Navy subjects, was closely homogeneous in both training and exposure to English.
5. Finally, one group of Vietnamese subjects was provided with both English text and translated text, so that the contribution to comprehension of a dual reading could be measured.

II. Method

Experimental Materials

The corpus of text from which the sample passage was drawn was selected for translation by the Naval Advisory Group, Vietnam. It presented 500 words from a larger U.S. Navy casualty control instruction. (See Appendix A.) It was judged by the Advisory Group to be representative of operational instructions and, as such, not easily read or understood in English. Flesch «Reading Ease» readability score (see Flesch, 1948) was 38, indicating that the material was approximately high school graduate or beginning college in reading level.* This level of difficulty, plus the technical nature of the material, would, we felt, prove to be an unusually strict test of a translator's skill.

Translations were produced by three methods. First, as a control condition we employed two highly skilled Vietnamese translators to provide manual translations. These men worked independently at first, then compared and modified their translations to produce a final consensus paper. Technical glossaries, insofar as they were available, were provided the translators. A professional Vietnamese linguist also reviewed this translation. These manual translations were produced using people and procedures likely to result in the highest quality of translation, a level not easily attainable in ordinary commercial or government facilities.

*) The Flesch scores for the three Air Force samples used in Sinaiko and Klare (1972) were, by comparison, 34, 39, and 25. Thus, the Navy sample material was about equivalent in difficulty to the two easiest Air Force passages.

Second, the LOGOS I system (see Byrne, 1970) was used to provide a rough un-edited computer translation. Although we sampled only 500 words from the larger English document, the LOGOS Development Corporation did not know which pages we could use. Third, the LOGOS Development Corporation also provided a post-edited version of the computer translation. Un-edited computer translations are considered to be misleading or even incorrect; therefore, MT is traditionally a two-stage process, the second of which is post-editing. (Both computer translations, un-edited and post-edited, were retyped in order to eliminate format as a variable.)

Appendix A contains the entire original English passage upon which our translations were based. The English material was also used as a control condition in two ways: (1) some Vietnamese Navy readers were tested on their comprehension of English, and (2) a group of U.S. Navy Officer Candidates served as readers.

All subjects were students at the Navy's Officer Candidate School (OCS) at Newport, R.I. The Vietnamese were all students in Classes 10, 11, and 12. The Vietnamese OCS program was essentially modeled after the American OCS, and since it is taught in English by U.S. Navy instructors, the Vietnamese Navy subjects were all qualified in English. Each class was tested within the last ten days of its 18-week course. The subject matter of the translation—casualty control—was generally familiar to the men but the particular document from which the translation was made had not been seen by them. The control group subjects consisted of 58 United States Navy Officer Candidates from Class 7109. These men were at about the same stage of training as the Vietnamese. And, like the VNN subjects, the USN subjects had had instruction in casualty control but none had seen the material we used.

All testing was done in a group situation. At each session there were from two to four test forms, some containing comprehension tests and some containing cloze forms, and test forms were distributed in a random order that ensured that adjacent men had different tests. The three major types of readability criteria were used (Klare, 1963); the specific measures used are described below.

- First, we constructed a four-alternative, multiple-choice test, with approximately one item for each of the 17 sentences in the sample passage. (See Appendix B for a copy of the English multiple-choice

test.) The rationale behind item construction was that items be deliberately easy; i.e., since we were measuring the comprehensibility of the text, rather than attempting to get maximum discrimination in knowledge among subjects, we wanted the reader to be able to earn a high score based solely on his reading. Put another way, we did not desire a distribution of test scores ranging from very high to very low. Further, subjects took the tests in an open-book mode, being able to refer to the text as often as necessary to answer test items. Test items were translated into Vietnamese for use with each of the three test forms in that language, i.e., human translation, un-edited computer translation, and post-edited computer translation. We also prepared cover sheets and sample test items. Scores consisted of the number of correct responses with no correction for guessing or wrong answers. In addition, subjects entered the time they began the reading tests and the time they completed the last item, referred to hereafter as «testing time.» Time for reading of the text was also separately recorded and is hereafter referred to as «reading speed.»

- Second, we built cloze procedure forms by systematically deleting each fifth word of text, starting with the second word. Subtitles were not included in our counting. This is standard procedure for preparing cloze forms and, since it has been elaborated elsewhere (Klare, Sinaiko, and Stolurow, 1971), it will not be further described here. Two scores were derived from the cloze tests: number of correct responses (misspellings were accepted, but not synonyms), and number of omitted or blank responses. We also prepared cover sheets explaining the use of the cloze procedure and including a practice sentence. Appendix C contains the cloze-formatted material in English. Subjects entered the times at which they began and finished cloze tests.

- We also used a newly revised intelligibility or clarity rating scale. The nine-point scale was adopted from Carroll (1966), and it had been successfully used in other experiments here (Sinaiko and Klare, 1972). The revised scale was used by subjects who had read the full text material, just before they answered the multiple-choice comprehension tests, and also by subjects who read the cloze forms, just after they filled in the blanks. A Vietnamese translation of the

scale was used with all material in that language. Appendix D contains the English form of the clarity rating scale.

The rationale for using several different indices is that they usually measure somewhat different aspects of readability. This is the case both conceptually and in terms of intercorrelations, as shown in our previous study (Sinaiko and Klare, 1972) and the literature on speed and comprehension. With this arrangement it is possible for the reader to select that index (or indices) of greatest importance to him.

Design

Each subject, selected randomly within his class, read and completed either the cloze-formatted material or the full text and the comprehension test. Vietnamese Navy subjects in the main experiment worked with one of the four possible language versions: English, human translation, unedited machine translation, or post-edited machine translation. United States Navy subjects worked only with English material. In the minor experiment, VNN subjects read and were tested, in both possible orders, on both the English text and the human translation. The two tests were alternated; i.e., if the multiple-choice test was given first, the cloze test followed, and vice versa. All results and discussion will pertain to the major experiment unless specifically labeled as a minor experiment.

III. Results*

Translation Mode

Table 1 summarizes the mean values, based on either 17 or 18 subjects each, for the three translation modes. Multiple-choice comprehension test scores and cloze accuracy scores are given as percentages correct, and cloze omissions scores as percentages of blanks, for ease of comparison.

*) All data have been subjected to rigorous statistical testing, using both analysis of variance and computation of significance of mean differences using Dunn's procedure (Kirk, 1968, pp. 79-81). We refer to «significant» results in the statistical use of the term; namely, that such observations can be considered highly reliable and not likely to occur by chance more than 1 in 100 times. In a few cases, differences are «less highly significant,» i.e., they would occur 1 in 20 times.
Full details concerning the statistical analysis of the data can be obtained directly from the authors.

Clarity ratings could range from a high of 9 (very clear) to a low of 1 (very unclear) on the scale.

TABLE 1. MEAN VALUES FOR EIGHT MEASURES OF READABILITY FOR THREE TRANSLATION MODES

Measure	*Translation*		
	Human	*Post-Edited MT*	*Un-edited MT*
Reading Comprehension, %	79	72	66
Testing Time, min	12.1	11.8	13.2
Reading Speed, min	5.8	5.6	6.8
Clarity Ratings	6.5	6.4	4.0
Cloze: Accuracy, %	55	41	27
Omissions, %	8	6	9
Clarity Ratings	6.4	4.8	3.2
Time, min	27	28	26

In terms of most of the measures of readability we used, human translation fared best of the three language versions, and un-edited MT fared poorest. This was the same rank order reported in the earlier IDA study (Sinaiko and Klare, 1972), suggesting that the effects are consistent. Mean reading comprehension (i.e., multiple-choice) test scores, for human, post-edited MT, and un-edited MT, were 79 percent, 72 percent, and 66 percent, respectively. Test-taking time was slower for the un-edited MT—13.2 minutes—than for either human—12.1 minutes—or post-edited MT—11.8 minutes. Reading speed was slightly slower, also, for un-edited MT—6.8 minutes—than for either human or post-edited MT—5.8 and 5.6 minutes, respectively. Clarity ratings of the three types of translation were in the predicted direction, i.e., human translation received the highest mean rating (6.5), edited MT was next (6.4), and un-edited MT was lowest (4.0); the latter rating was significantly lower than either of the other two.

Cloze accuracy score averages closely followed the same pattern as comprehension test scores: 55 percent, 41 percent, and 27 percent, respectively, for human translation, post-edited MT and un-edited MT. All differences were significant. The same was true for clarity ratings on cloze passages, which were 6.4, 4.8, and 3.2. Note that these ratings, in fact, followed the comprehension score pattern more closely than did the ratings on the regular text and the first of these was significantly

higher than the last. There was no discernible relationship between translation mode and omitted cloze responses—scores were low for all language versions: 8 percent, 6 percent, and 9 percent, suggesting that subjects did not find the cloze task a difficult one. Similarly, time to complete cloze forms was not closely related to translation mode, being 27.3 minutes, 27.9 minutes, and 25.9 minutes. Note, however, that un-edited MT close forms, on which accuracy scores were very low, were the fastest to be completed.

In summary, we believe that the readability of the Navy translations of material follows the same general pattern reported earlier in a similar experiment based on an Air Force text (Sinaiko and Klare, 1972). Translation by expert humans is most readable, followed by post-edited MT and un-edited MT. The cloze procedure appears to be the best single method of making the differentiation, although even a deliberately easy, multiple-choice comprehension test placed the three translation modes in the same order. Clarity ratings were also consistent in following this order, but the other, less-important time measures were not quite as consistent.

Comprehension of English versus translations

Table 2 summarizes the relative readability of English and Vietnamese translation versions for 17 VNN subjects on each version.

TABLE 2. MEAN VALUES OF EIGHT MEASURES OF READABILITY OF ENGLISH VERSUS VIETNAMESE: VIETNAMESE SUBJECTS

	Language	
Measure	*English*	*Vietnamese*
Reading Comprehension, %	80	79
Testing Time, min	13.9	12.1
Reading Speed, min	5.8	5.8
Clarity Ratings	6.1	6.5
Cloze: Accuracy, %	10	55
Omissions, %	56	11
Clarity Ratings	4.9	6.4
Time, min	25	27

Using the best version of Vietnamese (i.e., human) as a standard of comparison, the VNN subjects performed about equally well whether

they read English or the translation. This was true for such measures as reading rate, clarity ratings, and comprehension test scores. Scores on cloze procedure forms were dramatically different, favoring the subjects' primary language: accuracy of responses was 10 percent for English and 55 percent for the human-translated material. This difference was highly significant. The apparent paradox (i.e., almost identical performance on comprehension tests and very different accuracy levels on cloze) is easily explained. Cloze testing tends to emphasize the subtler aspects of language understanding (e.g., structural words over content words) while comprehension tests put a high premium on the reader's ability to understand content words (chiefly nouns). Thus, one would predict much poorer cloze scores for readers of a second language (who are taught to learn content words first) than for people reading their native tongue. This finding, incidentally, bears out an earlier study which arrived at the same conclusion (Sinaiko and Klare, 1972).

The ability of the VNN subjects to do as well on the English text as on the best of the translations (as far as reading rate, clarity ratings, and comprehension scores are concerned) suggests a further comment. Stated simply, perhaps the best way to help Vietnamese use U.S. manuals is to improve the readability of the English text itself. This could provide the considerable bonus of helping American users as well as users of other nationalities, and might be done at no more total cost (if as much as) than translation itself.

Reading performance of USN versus VNN subjects

Table 3 summarizes the relative readability of our material for 29 USN subjects (in English) and 17 VNN readers of the expert human translation.

Several interesting findings stand out in this comparison. First, USN subjects, reading 500 words of English text, were much faster than VNN subjects reading the same material in translation (this difference was significant). Yet reading comprehension scores were almost the same. Means for the two groups were: 2.7 minutes and 5.8 minutes for reading time, and 81 percent and 79 percent for comprehension, respectively.

TABLE 3. MEAN VALUES OF EIGHT MEASURES OF READABILITY OF ENGLISH VERSUS VIETNAMESE: USN AND VNN SUBJECTS

Measure	*Reader Group* *USN*	*VNN*
Reading Comprehension, %	81	79
Testing Time, min	5.2	12.1
Reading Speed, min	2.7	5.8
Clarity Ratings	5.5	6.5
Cloze: Accuracy, %	36	55
Omissions, %	6	11
Clarity Ratings	3.9	6.4
Time, min	15	27

Second, cloze procedure accuracy for USN subjects was much lower than was the case for VNNs: 36 percent and 55 percent, respectively (difference significant). At the same time, USN subjects spent less time filling in cloze forms (also significant). This suggests two explanations: (1) USNs showed a lower degree of motivation toward the tests; and (2) the human translators actually improved the quality of the original English text. Such an effect has sometimes been attributed to the translation process but no quantitative evidence exists to support this assumption. Furthermore, the difference in reading time for the English and Vietnamese noted above tends to support the motivation hypothesis, since both USN and VNN subjects were reading passages in their native language. And, finally, the difference in comprehension test-taking time—5.2 minutes for the USN group and 12.1 minutes for the VNN group (a significant difference)—also supports the motivation hypothesis. This suggests that the close correspondence in comprehension test scores noted above may simply not have been as meaningful as we first thought.

Air force and navy material compared

Our earlier experiment (Sinaiko and Klare, 1972) had a similar objective, and since it was based on a similar design it is possible to compare the readability of both the English and the three translation versions for the two studies. Both studies were based on 500-word samples of English material, three samples in the earlier study using USAF text and a single sample for the present study. Measures of readability were similar and, in some cases, identical. Both studies measured reading speed, both used the cloze procedure (from which were derived accuracy and omission

scores), and both studies measured time-to-complete cloze forms and time-to-complete reading comprehension tests. The measures differed in the following ways: (1) reading comprehension tests were of the fill-in type for the USAF study, and of the multiple-choice type for the present study; (2) a ninepoint clarity scale was used in both studies, but modified slightly for the second study in an attempt to make its descriptors and instructions more easily understood.

Table 4 summarizes the two groups of American controls, 29 USN Officer Candidates and 88 USAF student pilots, for six measures. The minor measures of testing time and ratings of cloze passages, presented in previous tables, are not reported here because of the difficulty of making a meaningful comparison. Also, because the data were obtained in two experimental settings, with some procedural differences between them, we chose not to do a statistical analysis.

Comparing both groups of American controls showed the following. Navy subjects tended to read faster than their USAF counterparts and their reading comprehension test average scores were slightly higher. However, the USN group spent somewhat less time on the cloze passages and mean cloze procedure accuracy scores were much lower for the USN group, i.e., 36 percent versus 56 percent. We suggest three hypotheses for this finding : (1) the readability of Navy material could have been poorer than the Air Force sample passages, despite a similarity in readability formula scores; (2) the Air Force material had been taken from a standard instructional text to which all of our USAF subjects had been exposed, while the Navy passage was from an operational instruction previously unseen by the USN readers; and (3) the USN subjects may have had a lower level of motivation toward participating in the study than the USAF subjects.

TABLE 4. MEAN VALUES OF SIX MEASURES OF READABILITY PERFORMANCE OF AMERICAN CONTROL GROUPS USN AND USAF

	Group	
Measure	*USN*	*USAF*
Reading Comprehension, %	81	77
Reading Speed, min	2.7	3.8
Clarity Ratings	5.5	7.7
Cloze: Accuracy, %	36	56
Omissions, %	6	3
Time, min	15	19

A reason supporting the hypothesis of greater difficulty in reading the Navy material is that clarity ratings were among the lowest of any we have observed, approaching those of un-edited MT. At the same time, USAF subjects' mean clarity rating was very high; respectively, average ratings were 7.7 (USAF) and 5.5 (USN). This indicates that the Navy material was perceived as more difficult. Also, we suggest that the Navy material, taken from an obsolete instruction first published about 1959, was less relevant to its readers than the USAF passages, which were directly related to the subjects' current activities. A reason supporting the third hypothesis is that, as noted above, other evidence points toward the low motivation hypothesis for USN subjects. The seeming contradiction that USN reading test scores were higher than USAF reading test scores could simply be due to the USN multiple-choice test being easier than the USAF fill-in test. This is generally the case in comparisons of these two types of tests. We cannot rule out the possibility, of course, that all three hypotheses might be partially true.

Readability of English for VNN and VNAF subjects

Table 5 summarizes the readability of the English text of the U.S. Navy material for 17 or 18 VNN subjects and the English text of the U.S. Air Force material for 12 to 14 VNAF subjects. (For the reasons given previously, statistical analyses were not run on these data.)

TABLE 5. MEAN VALUES OF SIX MEASURES OF READABILITY OF ENGLISH: VNN AND VNAF SUBJECTS

	Group	
Measure	*VNN*	*VNAF*
Reading Comprehension, %	80	69
Reading Speed, min	5.8	5.9
Clarity Ratings	6.1	6.5
Cloze: Accuracy, %	10	20
Omissions, %	56	48
Time, min	25	31

The readability of the English texts, on naval material for VNN subjects and on air force material for VNAF subjects, provides several tentative but useful comparisons. First, the VNN subjects received higher average scores than the VNAF subjects on reading comprehen-

sion, but lower average scores on the cloze test. This suggests again that the multiple-choice test taken by the VNN group may well have been somewhat easier than the fill-in test taken by the VNAF group, a possibility we raised earlier.

Second, VNN and VNAF subjects spent almost equal time, on the average, reading their respective English texts. The naval material in English appeared to be less readable, on casual observation, than the air force material in English, which would normally suggest longer reading time for the former. The readability formula scores reported earlier, however, failed to show much difference between the naval and air force material, which suggests that the difference in readability of the two may well have been more apparent than real.

Results: Minor experiment

We indicated earlier that several groups of VNN subjects were given both English text and the corresponding human translation, i.e., two versions, with appropriate comprehension test or cloze format, to see if this arrangement improved comprehension when compared with groups having only a single version. This dual presentation permitted search of the corresponding version when difficulties in comprehension arose, an arrangement sometimes considered to be of value in difficult translated material. Our study was admittedly a crude test of this hypothesis, since one of the versions was, in all cases, a cloze-formatted version, which is likely to be of less help than a full-text version.

Table 6 presents data for all of the groups in the minor experiment, based on from 15 to 21 subjects for the various groups.

We found the following, which we believe indicative despite the above reservation. First, cloze accuracy scores on the human version were helped little, if any, by providing an English version first. The group (designated B in Table 6) with such a corresponding version received an average score of 57 percent, while the group which had only the human version had an average score of 55 percent.

Second, scores on the English cloze version were helped considerably if they were preceded by a human translation. The group with both versions (A) had a cloze score of 29 percent while that with the cloze version only (G) had a score of 10 percent. This difference was significant. The cloze test in English, a second language for the VNN group, has been found very difficult consistently in our work; in such a case, having

a translation is clearly helpful in searching for answers. The extent of the searching is shown by the fact that subjects took only 25 minutes to complete the cloze test if they had no preceding human translation (Group G), whereas they took 45 minutes on the cloze test if they had one (Group A). This difference is significant. But even a translation can provide only limited help: the English cloze score of 29 percent is, even with the text of the human translation available, only about half of what the cloze score is on the human translation itself (i.e., around 55 percent, as in Group F).

TABLE 6. MEAN VALUES FOR EIGHT MEASURES OF READABILITY FOR THREE GROUPS HAVING DOUBLE VERSIONS AND FOUR GROUPS HAVING SINGLE VERSIONS OF EXPERIMENTAL MATERIAL

Group A First [1]) Version *Human*	*Group B* First Version *English*	*Group C* Second Version *English*	*Group D* Only Version *Human*	*Group E* Only Version *English*
ıding				
:omprehension, % 79	78	67	79	80
ting Time, min 11.1	10.9	13.4	12.1	13.9
ıding Speed, min 5.0	5.9	8.4	5.8	5.8
rity Ratings 6.4	6.4	5.7	6.5	6.1

	Group A Second[1]) Version *English*	*Group B* Second Version *Human*	*Group C* First Version *Human*	*Group F* Only Version *Human*	*Group G* Only Version *English*
ze: Accuracy, %	29	57	53	55	10
Clarity Ratings	4.4	6.3	5.1	6.4	4.9
Omissions, %	20	5	7	8	56
Time, min	45	32	36	27	25

Finally, having a human cloze version was of no help for a reading test on a following (second) English version; in fact, scores were lower (67 percent versus 80 percent) when the human cloze version was present (Group C) than when it was absent (Group E). This difference was significant. We have no ready explanation for this except some possible interference on a difference in the groups.

[1]) First and second refer to order of presentation to a particular group of subjects.

We did not have a group with an English cloze version followed by a human translation and comprehension test (which would have been the fourth possible group with two versions) for two reasons:

1. The number of subjects available was limited, and having another group would have cut the size of each of the groups too much.
2. We have found that a few VNN subjects are likely to work so long on an English cloze test, that when it is presented first, it may interfere with subsequent activities. This group, therefore, seemed the best one to eliminate.

Correlations between measures

As indicated earlier, eight measures of readability were used: reading comprehension (multiple-choice) test; test-taking time; reading speed; clarity rating; cloze accuracy score; cloze omission score; cloze clarity ratings; and cloze time. In such a situation, there is some question of whether all of the measures are needed. If scores on one measure correlate highly enough with scores on another, only one of them (presumably the more accurate, reliable and/or easier) should be sufficient in the future. If, however, the correlation is low, the two are measuring different aspects of readability and both may well have a place in future testing.

Since each type of translation was tested with different groups of subjects, a total of 60 correlation coefficients were run on the groups in the major experiment (similar correlations for the minor experiment would not be as readily interpretable). Of these, 30 involved the several cloze-format measures and 30 the several comprehension-test-format measures. The most clear-cut relationship was that between cloze accuracy score and cloze omissions, as had been the case in Sinaiko and Klare (1972). The correlations ranged from —0.77 to —0.39, not quite as high as previously, but all statistically significant. This suggests that the number of omissions might not as readily serve to provide estimates of cloze accuracy as previously, but might be used as an estimate or on a «last resort» basis.

In contrast to the results reported in Sinaiko and Klare (1972), certain correlations involving the clarity ratings were significant in this study. Three of five correlations involving cloze scores and cloze ratings were statistically significant, as were three of five involving reading time (before taking the comprehension test) and clarity ratings. It seems

likely that the improvements made in the rating scale since its use in Sinaiko and Klare (1972) were responsible for the above results. In general, it seems desirable to further use the rating scale experimentally to establish its value in translation work.

IV. Discussion

Translation mode

There is little doubt that translation by expert human linguists produces a more readable document than the best machine translation. This is particularly true when the most stringent means of measuring language comprehension is used, i.e., cloze procedure. Of the two types of MT used in the experiment—un-edited and post-edited—the latter was always better understood as measured by both comprehension tests and cloze. However, although post-edited computer translations may approach good human translation for readability level, they still have a long way to go. Un-edited MT, except perhaps for the least technical material we studied, is probably not worth presenting. We have shown earlier (Sinaiko and Klare, 1972) that about forty weeks and $ 1,000 are required to train a single Vietnamese to the level of comprehension in English that permits instruction in that language. The present study confirms our earlier findings that English can be understood by Vietnamese who have a modicum of exposure to the language. However, the subtler aspect of the language (e.g., idiom and structural terms) is not as well handled by non-native readers. Because of the apparent sensitivity of the cloze procedure to comprehensibility of a second language, we suggest that the technique might be useful to teachers as a measure of growth in language skill. Thus, teachers of English-as-a-second-language, as well as teachers of a foreign language to Americans, could use cloze as a convenient indicator of the students' grasp of the most difficult aspects of the languages they are learning. We have not seen this use of the cloze procedure before.

Methodological consideration

Our earlier work with VNAF subjects reported our concern about the meaningfulness of a rating scale technique when used by Vietnamese subjects. It is our observation, based on large discrepancies among ratings

and other more objective performance measures, that many of the subjects simply did not understand the use of ratings. As a result of that experience, we re-wrote the instructions for our clarity rating scale as well as some of the nine alternative rating descriptors. Vietnamese Navy subjects, using the revised scale, did not appear to experience nearly the difficulty their VNAF counterparts had with the older version, and their ratings were more consistent with other measures. Therefore, we recommend that consideration be given to much wider use of the Vietnamese version of our scale (see Appendix D) as a tentative means of indicating relative readability in the field.

When we proposed using multiple-choice-type test items our expert translators felt that such a format would present difficult, if not impossible, obstacles to translation. As seen by the accuracy scores for the reading comprehension tests, such was not the case. Since multiple-choice items are much easier for the researcher to handle—they do not require a native reader of Vietnamese for scoring—we recommend the continued use of this type of item. (The superior performance of the VNN subjects with multiple choice might have been a reflection of their experience with this form in training. We do not know, however, that this is actually the case.)

Cloze, although a severe test of readability, discriminates different quality levels of translations, and can be prepared and scored very easily. This is somewhat less true for the clarity scale although very poor translations—e.g., un-edited MT—always show the lowest ratings. Reading comprehension tests, either multiple-choice or fill-in format, are also good discriminators of translation quality. Multiple-choice tests, of course, have the advantage that they are more easily and objectively scored than fill-in items. Reading speed, at least as we estimated it from self-timed tests, was of some value in discriminating versions, with faster reading generally occurring with what we judged to be the better version of English. (This was true for translated material, not for untranslated English.)

Finally, administration of both an English text and a translation appears to help on a very difficult task like completing cloze blanks in a second language. It appears to be of doubtful value in an easier task such as completing cloze blanks in a subject's native language or even taking a fairly easy reading comprehension test in a second language.

REFERENCES

BYRNE, C.E., B.E. SCOTT, and T.N. BINH, «Demonstration of LOGOS I System for English-Vietnamese Machine Translation,» Rome Air Development Center, Griffiss Air Force Base, New York, RADC-TR-70-170, August 1970.

CARROLL, J.B., «An Experiment in Evaluating the Quality of Translations,» Appendix 10 in Publication 1416, *Language and Machines: Computers in Translation and Linguistics.* Washington, D.C.: National Academy of Sciences—National Research Council, 1966.

FLESCH, R.F., «A New Readability Yardstick,» *Journal of Applied Psychology, 32,* 221-233, 1948.

KIRK, R.E., *Experimental Design: Procedures for the Behavioral Sciences,* Belmont, California: Brooks-Cole, pp. 79-81, 1968.

KLARE, G.R., *The Measurement of Readability,* Ames, Iowa: The Iowa State University Press, 1963.

KLARE, G.R., H.W. SINAIKO, and L.M. STOLUROW, The Cloze Procedure: A Convenient Readability Test for Training Materials and Translations,» IDA Paper P-660, Institute for Defense Analyses, Arlington, Virginia, 1971.

SINAIKO, H.W., and R.W. BRISLIN, «Experiments in Language Translation: Technical English-to-Vietnamese,» Research Paper P-634, Institute for Defense Analyses, Arlington, Virginia, July 1970.

SINAIKO, H.W., and G.R. KLARE, «Further Experiments in Language Translation: Readability of Computer Translations,» *I.T.L., 15,* 1-29, 1972.

APPENDIX A: «CASUALTY CONTROL» TEXT

4. *Lubricating Oil Service Systems.* The lubricating oil service systems for the main propulsion units comprise six separate groups, one for each of the four main diesel engines and one for each reduction gear-clutch unit. Each system is designed for independent operation and no interconnection is provided for normal service. The lubricating oil service systems for the ship's service and emergency diesel-generator sets are integral with these units.

a. *Main Engine Lubrication.* Each main engine is equipped with a pressure lubrication and piston cooling system which supplied a continuous flow of oil to all surfaces requiring lubrication, and to the pistons for cooling. The system is served by a positive displacement gear-type lubricating oil pump, driven from the lower crankshaft. Since reversal of the engines for astern operation also results in reversal of rotation of the attached pumps, a reversing valve is provided for each engine, which interchanges the suction and discharge connections of the corresponding attached lubricating oil pump simultaneously with the change of main engine rotation.

(1) *Pressure System.* The pump takes suction from its sump tank through the strainer box, swing-check valve and reversing valve, and discharges at a normal pressure of approximately 40 p.s.i. through reversing valve, strainer, filter and

lubricating oil cooler to the service headers which furnish the internal lubrication and piston cooling oil supply for the engine. A relief valve in the pump discharge, set at 60 p.s.i., protects the pump and piping against excess pressure. Lubricating oil returns from the engine drain internally to the sump tank. Each lubricating oil filter is vented to the corresponding sump tank and is fitted with a drain to the bilge.

(2) *Temperature Control.* An automatic temperature regulating valve, controlling the flow of seawater through the lubricating oil cooler, is provided for maintaining a constant lubricating oil temperature at the cooler outlet. The normal recommended temperature range is 120° to 140° F. In case of failure of the automatic system, the regulating valve may be adjusted manually to any desired valve position. The oil pressure should always be higher than the water pressure so that if a leak should develop in the cooler, water would be prevented from entering the oil system. The oil cooler is provided with a relief valve bypass, set to function if the pressure drop through the cooler exceeds 20 p.s.i. The arrangement is designed to protect the cooler from excess pressure due to clogging or under cold starting conditions, and also to assure a supply of oil to the engine under such circumstances.

(3) *Alarm.* A low pressure alarm, set to sound if the oil pressure to the engine falls to 8 p.s.i. is installed in the system at the oil inlet to the engine. The alarm contactor is provided with an idling output to make the alarm inoperative while the engine is being reversed, or when at speeds below 200 «R.P.M.» A control interlock is arranged to stop the engine if the pressure drops below 5 p.s.i.

Appendix B: Multiple choice test (English)

Instructions

This is a simple test of your understanding of the material you just read. There is a series of multiple-choice questions. Each question has only one correct answer. You should draw a circle around the letter in front of the correct answer.

You may refer back to the text if you wish. Be sure to write in the time you start and finish the test.

Sample question. The U.S. Navy Officer Candidate School is located in:

a. Washington
b. Norfolk
c. Chicago
d. Newport

The correct answer is «d. Newport» and you should have circled that letter.

1. How many lubrication oil service systems are there?

   a. 1
   b. 2
   c. 4
   d. 6

2. During normal service of the lubricating oil systems the following kind of operation is used:

   a. interconnected operation
   b. independent operation
   c. both interconnected and independent operation
   d. sometimes interconnected and sometimes independent

4. Pressure lubrication and piston cooling systems

   a. supply a continuous flow of oil to surfaces that need it
   b. are provided for only one of the main engines
   c. supply a flow of oil to each piston in sequence
   d. are served by the emergency generator

5. The lubricating oil pump of the main engine cooling system is driven

   a. by a gear-type piston
   b. by a pressure pump
   c. from the lower crankshaft
   d. from a gear-type pump

6. Reversing the engines causes

   a. suction and discharge to stop
   b. no change in rotation of the pumps
   c. a delay in direction of rotation
   d. reversing of the attached pumps

7. The main engine lubricating oil pump takes suction from

   a. its reversing valve
   b. its sump tank
   c. the lubricating oil cooler
   d. the service headers

8. The setting for the relief valve in the pump discharge is

   a. 60 p.s.i.
   b. 40 p.s.i.
   c. 20 p.s.i.
   d. 8 p.s.i.

9, 10. Each lubricating oil filter is vented

a. to its own sump tank
b. to the bilge
c. to the engine drain
d. to the swing check valve

11. Constant lubricating oil temperature is maintained by

a. a flow of seawater to the bilge
b. the lubricating oil cooler
c. a cooler outlet
d. an automatic temperature regulating valve

12. The normal temperature of the oil at the oil cooler outlet

a. must remain exactly constant
b. can vary over a range of 120° to 140° F.
c. should go no higher than 120° F.
d. should exceed 120° F

13. If the automatic temperature regulating system fails

a. the valve may be adjusted manually
b. a standby system takes over
c. main engines should be shut down
d. oil temperature rises very slowly

14. Water is prevented from entering the oil system because

a. leaks cannot occur
b. oil pressure in the cooler is kept higher than water pressure
c. water pressure is higher than the oil pressure
d. of a regulating valve system

15. The oil cooler has a relief bypass valve which operates

a. if the pressure rises more than 40 p.s.i.
b. when the pressure drop exceeds 20 p.s.i.
c. intermittently only
d. by manual control

16. The purpose of the relief bypass valve is

a. to protect the cooler from excess pressure
b. to prevent clogging under warm weather or tropical conditions
c. to control clogging
d. to maintain oil pressure at 60 p.s.i.

17. A low pressure alarm is set to sound if

    a. oil inlet pressure drops to 1 p.s.i.
    b. there is a water leak into the lubricating system
    c. pressure drops to 8 p.s.i.
    d. there is any pressure change in the cooler system

18. The low pressure alarm does *not* sound if

    a. engines are being operated at speeds over 300 rpm
    b. the engine is being reversed
    c. the automatic cutoff system is working
    d. temperatures are between 120° and 140° F

19. A control interlock system stops the engine when oil pressure drops below

    a. 1 p.s.i.
    b. 2 p.s.i.
    c. 5 p.s.i.
    d. 10 p.s.i.

Appendix C: Cloze format (English)

Instructions: Cloze Test

This is a new type of test. It was made by copying a part of a Navy instruction and leaving out every fifth word. You are to think of the correct word for each blank and write in it the proper space. Only one word goes in each blank. Guess if you are not sure. Write carefully. Try this sample sentence:

With the fuel _____ under prime, attempt to _____ _____ the engine. If the _____ will start but there _____ no fuel pressure when _____ is secured, replace the _____ pump. The engine may _____ _____ operated on priming fuel _____ until the fuel pump _____ be replaced.

The correct words you should have filled in are: system, start, engine, is, priming, fuel, be, pressure, can.

4. *Lubricating Oil Service Systems.* The _____ oil service systems for _____ main propulsion units comprise _____ _____ separate groups, one for _____ of the four main _____ engines and one for _____ reduction gear-clutch unit. _____ system is designed for

_____ operation and no interconnection _____ provided for normal service. _____ lubricating oil service systems _____ the ship's service and _____ diesel-generator sets are _____ with these units.

a. *Main Engine Lubrication.* Each _____ engine is equipped with _____ pressure lubrication and piston _____ _____ system which supplies a _____ flow of oil to _____ _____ surfaces requiring lubrication, and _____ the pistons for cooling. _____ system is served by _____ _____ positive displacement gear-type _____ oil pump, driven from _____ lower crankshaft. Since reversal _____ _____ the engines for astern _____ also results in reversal _____ rotation of the attached _____, a reversing valve is _____ for each engine, which _____ _____ the suction and discharge _____ of the corresponding attached _____oil pump simultaneously with _____ change of main engine _____.

(1) *Pressure System.* The pump takes suction _____ its sump tank through _____ strainer box, swing-check _____ _____ and reversing valve, and _____ at a normal pressure _____ approximately 40 p.s.i. through _____ _____ valve, strainer, filter and _____ oil cooler to the _____ headers which furnish the lubrication and piston cooling _____ supply for the engine. _____ relief valve in the _____ discharge, set at 60 _____ _____, protects the pump and _____ against excess pressure. Lubricating oil filter _____ vented to the corresponding _____ tank and is fitted _____ a drain to the _____.

(2) *Temperature Control.* An automatic temperature regulating _____ _____, controlling the flow of _____ through the lubricating oil _____, is provided for maintaining _____ _____ constant lubricating oil temperature _____ the cooler outlet. The _____ recommended temperature range is _____ to 140° F. In case _____ failure of the automatic _____, the regulating valve may _____ _____ adjusted manually to any _____ valve position. The oil _____ should always be higher _____ the water pressure so _____ if a leak should _____ _____ in the cooler, water _____ be prevented from entering _____ oil system. The oil _____ is provided with a _____ valve bypass, set to _____ _____ if the pressure drop _____ the cooler exceeds 20

_____. This arrangement is designed _____ protect the cooler from _____ pressure due to clogging _____ under cold starting conditions, _____ also to assure a _____ of oil to the _____ under such circumstances.

(3) *Alarm.* A _____ pressure alarm, set to _____ _____ if the oil pressure _____ the engine falls to _____ _____ p.s.i. is installed in _____ system at the oil _____ to the engine. The _____ contactor is to make _____ alarm inoperative while the _____ _____ is being reversed, or _____ at speeds below 200 _____. A control interlock is _____ to stop the engine _____ the pressure drops below _____ _____ p.s.i.

Appendix D: Clarity rating scale (English)

Scale of Clarity
Instructions

Some passages are easier to understand than others. We need your help in judging the clarity of the passage you have just read.

First, please read the nine numbered statements below. Second, decide which one of them describes the passage you have read. Third, put an «X» only in the box beside that one description.

9 — Perfectly clear and easy to understand. Says things very well.
8 — Almost perfectly clear and quite easy to understand. Could be easily changed to say things better.
7 — Generally clear. However, the words or the sentences used are not as helpful to a reader as they could be.
6 — General idea becomes clear very quickly. However, full understanding comes slowly to a reader because of the words or the sentences used, and the way of saying things.
5 — Clear only after considerable study. A reader can then be fairly sure he understands the main idea, even though the words or the sentences used make this harder than it needs to be.
4 — Seems pretty good, but is really quite hard to understand. The idea is not clear to a reader, and the words or the sentences used are poor.
3 — Generally *not* clear. Does *not* say things well, but after considerable study a reader can at least guess what the main idea is.
2 — Can hardly be understood at all. After considerable study, however, a reader feels that there is some kind of main idea.
1 — Cannot be understood at all. No amount of study would help a reader know what the main idea is.