[*ITL (Review of Institute of Applied Linguistics)* 15, 1972, 1-29]

# FURTHER EXPERIMENTS IN LANGUAGE TRANSLATION: READABILITY OF COMPUTER TRANSLATIONS

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

George R. KLARE
*Ohio University*

## ABSTRACT

Application of computational linguistics, i.e., language translation by computer, has been proposed as a means of producing readable translations of technical English-to-Vietnamese.

This report is about an experimental study of the readability of translations that could be used for training or equipment maintenance.

The experiments involved assessing the readability of Vietnamese that had been translated from English by three methods: (1) expert human translators, (2) un-edited translation by computer, and (3) edited computer translation. English was a control condition. Readers included two groups of student pilots: 168 in the Vietnamese Air Force (VNAF) and 88 in the USAF. Material that was translated consisted of three 500-word passages sampled from a standard Air Force text, *Instrument Flying.* Readability was measured by : (1) reading comprehension tests, (2) cloze procedure, and (3) clarity ratings. Time to complete each of these tasks was also measured.

Major conclusions of the study are : (1) expert human translators produce more readable translations of technical English-to-Vietnamese than is done by computer; (2) Vietnamese readers, trained in English, show the highest comprehension when dealing with that language; (3) comprehension *loss* becomes relatively greater, as more and more difficult material is read, for computer-based translations than for human translations; (4) method of translation docs not affect reading speed.

## I. Background

### *The problem*

The ability to read and understand technical material is a critical skill in learning to operate and maintain complex equipment. Textbooks and manuals are at the heart of training programs and technical orders; field manuals and other maintenance documents provide information used by the trained technician. Where military assistance or foreign aid programs are involved in transferring equipment to other nations, the language of the documents is particularly important; either the reader must learn English or the material must be translated. Put another way, the ability to deal with a foreign language becomes central to the larger technology transfer problem; either the recipient works in a second language or the original document is translated. Both approaches present formidable problems.

The translation of technical material is specially vulnerable to error. Moreover, meaning changes in translated documents can render the technician's task impossible; if he cannot understand a manual (or worse, if he reads a manual erroneously translated), either the maintenance of equipment cannot be done or it is done incorrectly. An earlier IDA (Institute for Defense Analyses) study (Sinaiko and Brislin, 1970) demonstrated experimentally that the quality of translated material is directly reflected in the accuracy with which technicians can do their work. The same study also showed that good translations from English to Vietnamese resulted in the same high quality of work by VNAF technicians that U.S. Army mechanics produced using English. Poor translations into Vietnamese, however, caused more maintenance errors than was the case for those men reading a second language (i.e., Vietnamese technicians using English).

Apparently, then, the optimal solution is to provide people with text that has been translated with high fidelity into their own native language. But translation, for any pair of original and target languages, is costly; good translations are produced slowly and good technical translators are in short supply. The best evidence we have about productivity of professional translators points to about 3000 words a day; interestingly, this seems to hold for several language pairs, e.g., Russian-to-

English, English-to-Vietnamese, French-to-English (Pierce, 1966). There is another, more serious drawback to relying on linguist-translators. For the type of high-quality work demanded by technical material only a small proportion of bilinguals are competent as translators. Technical translation demands not only skill in the two languages but some knowledge of the subject being translated. Ideally, a good technical translator would be grounded in the subject matter he is treating. Practically, however, there are few translators with relevant training (or few technical people who are also qualified as translators). In practice, good translation systems try to provide aids—glossaries, technical word lists and, if possible, the services of a technical person to explain concepts to translators.

Vietnamese presents a special problem as the target language of technical translation: there are very few technical terms in the language (although it is lexically very rich otherwise). Translators having to deal with technical material must resort to any of several means to handle the language: they may coin words, they may transliterate English, or they may describe the English term in functional Vietnamese, if possible. Thus, the term «tachometer» may become «rotation measuring machine» in Vietnamese (Sinaiko and Brislin, 1970). However the translator chooses to handle such terms, the task is made extremely difficult and much slower than non-technical translation. Further, there are no standard terms available to the translator. Thus, any means that can aid the linguist-translator would be most useful.

## *Machine translation*

For about a decade and a half there have been a number of research and development attempts to exploit computers in the language translation process. In their ultimate application computers would handle the entire task from input in the original language to finished output in the target language. Most of the machine translation (MT) effort has been in the United States[1]*)* and nearly all of this has concentrated on Russian-

[1]) The Department of Defense, chiefly through Air Force sponsorship, alone spent nearly $12 million during 1953-65, the last year for which accurate figures are available. During approximately the same period expenditures on MT in the CIA (Central Intelligence Agency) and National Science Foundation were $1.3 and $6.5 million, respectively (Pierce et al., 1966).

to-English. MT has, unfortunately, had a generally disappointing history. Early proponents of the notion of computer-mediated translation vastly underestimated the structural complexities of the language pairs involved; in fact, some of the first proponents of MT assumed that what was needed amounted only to a high-speed dictionary look-up procedure and computers would be ideally suited for that purpose. The promise of MT has only recently begun to be realized for Russian-to-English on a production basis and, at best, the system currently in use at the Foreign Technology Division, USAF, still requires considerable post-editing by skilled linguists. Post-editing is the process of correcting rough computer translations so that they are accurate and intelligible; editors who do this type of work are themselves bilingual and they add a substantial cost increment to the overall translation process. Estimates of the cost of post-editing, in the total MT process, have gone as high as 70 percent (Anonymous, 1965). Furthermore, the task of post-editing demands the same bilingual skills as conventional translating (Pierce et al., 1966).

In spite of its obvious handicaps, MT still holds great promise for translation of large volumes of material. A recent list of U.S. Army field and technical manuals awaiting translating in Vietnam numbers over 800 titles. English-to-Vietnamese has provided a particularly difficult translation problem, in addition to the absence of technical terms, because of a shortage of qualified translators. This further underscores the potential value of an effective MT system.

During the last year or two an English-to-Vietnamese MT system has been in development under Air Force sponsorship. This is known as the LOGOS I system (after the name of the LOGOS Development Corporation, contractor to the Air Force Systems Command). LOGOS I was demonstrated publicly in June 1970 and its output was judged sufficiently promising to warrant a more complete trial (Byrne et al., 1970). In the fall of 1970, the Training Directorate, U.S. Military Assistance Command, Vietnam, selected and forwarded for translation several technical manuals representative of the type of material that Vietnamese trainees and technicians would use. One of the documents, Air Force Manual 51-37, *Instrument Flying,* eventually was put through the LOGOS I system and it became the corpus of the present study.

There is great interest in an MT system that can handle technical English-to-Vietnamese, not only for Air Force use but across all Service needs. As part of its task to study some problems of Vietnamization for the Office of the Director, Defense Research and Engineering, Deputy for Southeast Asia Matters, IDA undertook an experimental assessment of the translations produced by LOGOS I. This paper reports this work.

*This study*

Our objective was to answer these questions:

- Given technical English that has been translated into Vietnamese by computer, what is the readability or comprehensibility of such translations?
- How do they compare with the same material that has been translated by highly skilled linguists?
- What is gained, in readability, in the post-editing process, i.e., between initial rough computer output and final, edited text?
- How does readability of translations into Vietnamese, done either by machine or by human, compare with that of the original English?
- How do Vietnamese readers fare when handling English as a second language versus reading the material after it has been translated, either by computer or linguist?
- How do Vietnamese readers, using either English or translations into Vietnamese, fare compared to American readers using English text?

There are many other questions that can, and should, be directed to the MT process :

- How much would a production MT system cost?
- How much lime and effort would be involved in preparing material for MT processing?
- What are the true costs, in time and money and skilled linguists, of post-editing?
- How can these costs be reduced?
- What is the availability of translators and post-editors in the U.S. and Vietnam?
- Is the need for translation sufficiently urgent to justify relatively imperfect translations?

These are important issues and they should be addressed. However, they were not part of the present inquiry.

### *Recapitulation*

This paper reports an experimental study of the readability of technical material that has been translated into Vietnamese. The translation mode took three forms: (1) linguist-mediated or human translations, (2) rough, or un-edited, output of a developmental MT system, and (3) finished, or post-edited, MT translations. For experimental control purposes we measured the readability of the original, untranslated English as read by Vietnamese and by USAF personnel. Our main interest was in the human variables of comprehension, rate of work, and judgments of clarity of the material. We addressed cost factors of the translation processes only peripherally. We did not look into the many linguistic aspects of the translation modes,

## II. METHOD

### *Reading material*

The experimental corpus consisted of three passages from Air Force Manual 51-37 (USAF, 1968), *Instrument Flying.* Each passage contained approximately 500 words,[2]) in order to get a good test of readability. They were selected to represent material of different levels of technical complexity and complexity was estimated by the experimenters, as described more fully below. Another reason for such a selection was the possibility that material from different chapters might well differ in readability. No other qualifications were placed upon the selections (i.e., they were randomly selected within the above restriction). The sample passages finally selected are described below.

1. Chapter 1, «Evolution of Instrument Flying». A total of 530 words from the section, «Early Flight Instrumentation», Flesch «Reading Ease» score (Flesch, 1948)[3]) for the passage was 34, or approximately

[2]) Note that the word-counts below are based on the method of counting hyphenated words as single words, which is done for many readability counts. Words in headings have also been included in the counts for completeness, though such words are not usually included in readability counts.

[3]) Reading ease is derived from a formula that takes into account average sentence length in words and average word length in syllables. Resulting scores range from 0 (practically unreadable) to 100 (easy for any literate person). For detailed instructions on the use of the formula see Flesch (1948) and Klare (1963).

high school or beginning college in reading level. This material appeared to be a straight prose passage, the least technical of our samples.

2. Chapter 3, «Differential Pressure Instruments». A total of 533 words from the section, «The Vertical Velocity Indicator», through the beginning of the next section, «The Airspeed Indicator». Flesch score for the passage was 39, giving approximate reading ease of high school or beginning college level.
3. Chapter 16, «Instrument Landing System». A total of 554 words from the beginning of the chapter and into the section, «Equipment and Operation». Flesch score was 25, indicating a reading level of college graduate. This sample appeared the most highly technical of the three we chose.

In addition to the original in English from the manual, each passage was also available in three translations into Vietnamese; however, the few illustrations present in the manual were removed for this experiment because our concern lay in the language rather than a mixed presentation of words and pictures. The experimental versions of each passage were the following.

1. English-language versions taken directly from AFM 51-37, *Instrument Flying.*
2. Vietnamese translated by a team of two of the most expert translators available. They used essentially the same method that had been shown, in an earlier experiment (Sinaiko and Brislin, 1970), to provide very high-quality translations: the men worked independently at first, then they reviewed and criticized the other's translations, then they rewrote a «consensus» version, which was reviewed by a Vietnamese linguist-consultant. This is to emphasize that we used translators and procedures that provided as high a quality translation as the experimenters had available. The method was, of course, a costly one and this was done in order to provide a standard of excellence against which to judge other translations. The best available technical English-Vietnamese glossaries were provided to the translators.
3. Vietnamese translated by computer, i.e., the LOGOS I System (Byrne et al., December 1970), and without benefit of further editing.

4. The same Vietnamese translated by computer, but additionally subjected to a thorough post-editing process. This was done by a team of bilingual Vietnamese who worked at the LOGOS Corporation and who used the original English material,

It should be noted that both computer translations (un-edited and post-edited) had been prepared as part of a task involving the translation of all of AFM 51-37. Thus, the passages used in this experiment were chosen randomly and independently by the authors from the available material. The LOGOS Corporation, which provided good cooperation throughout our study, did not know which particular chapters or passages were to be used in the IDA study,

All versions, both English and Vietnamese, were retyped so that format was not a variable.

*Subjects*

Two groups of readers were tested. The first was a sample of 88 USAF student pilots at Craig Air Force Base, Alabama. All were college graduates; approximately half had entered training in April 1970, and half in July 1970. All had used Air Force Manual 51-37 to some extent prior to testing. All subjects read English versions of the experimental materials, with the purpose of the testing being primarily to provide a base for comparison with the Vietnamese subjects who were tested subsequently.

The second group consisted of 172 VNAF student pilots in training at Keesler Air Force Base, Mississippi. (The data from 4 of the 172 subjects tested during the morning of March 30 had to be discarded owing to deficiencies in the test booklets, so that data from 168 subjects were available for analysis.) All of the VNAF subjects had at least a «first baccalaureate» (high school equivalent) education, and some had been to college. The VNAF student pilots had studied at least two foreign languages, usually French and English, before entering the military service. Prior to leaving Vietnam, and after basic training, all had been through the English language course supported by the Defense Language Institute. All students had reached a minimal English Comprehension Level (ECL) of 70 before going to the United States for training. Upon arrival they had been sent to Lackland AFB for 15 weeks of additional English instruction, specializing in aviation-related terminology. Although all VNAF students had been issued a copy of AFM 51-37, it was not possible to determine the extent to which each man had read the volume. As will be shown below, the VNAF subjects represented a wide range of training experience, a fact over which we had no control and which could have influenced test performance. However, the test administration procedures ensured that most experimental conditions were distributed across all classes.

The VNAF subjects began their training at different times, so their exposure to the material in the tests was quite varied. This can be seen most readily from the fact that the length of stay (i.e., weeks in training), at Keesler AFB varied from 3 to 27 weeks at the time of our testing. This, in addition, also influenced the men's command of English. A summary of the classes from which the subjects came shows the following:

| Class | Number of Subjects | Weeks of Training |
|---|---|---|
| 71-07 | 30 | 27 |
| 71-08 | 32 | 22 |
| 72-01 | 34 | 15 |
| 72-02 | 40 | 9 |
| 72-03 | 32 | 3 |

Note, finally, that the first three classes (those who had been at Keesler from 15 to 27 weeks) had been through an instrument procedures course, while the last two classes had not. However, all students had been exposed to a two hour introduction to radio tuning procedures.

## *Test procedure*

The procedure will be described in terms of independent (or main) experimental variables, dependent variables (or measures), experimental design, and administration.

### 1. *Independent Variables*

Of greatest interest was comparison of the readability of the several versions of the Vietnamese translation with each other and with the English text. Of related interest was comparison of the relative readability of the three samples—the two technical passages and the non-technical passage. The two independent variables used and their abbreviations were:

a. *Versions or Language Conditions*

1) English (EN)
2) Vietnamese translation by humans (HU)
3) Vietnamese translation by computer, post-edited (PE)
4) Vietnamese translation by computer, un-edited (UE)

b. *Passages*

1) About 500 words from Chapter 1 (C1)
2) About 500 words from Chapter 3 (C3)
3) About 500 words from Chapter 16 (C16)

## 2. *Measures*

We used the three major types of readability criteria: comprehension, rate of work, and judgment of the acceptability (clarity) of the material (see Klare, 1963). Each is described below.

a. *Comprehension.* Comprehension was measured in two ways. The first was a reading test consisting of completion or fill-in items. Subjects worked in an «open-book» mode; i.e., they could refer back to the text if they wished. Because of the nature of the material, there were maximum possible scores of 13 on the reading tests for Chapters 1 and 3, and 16 for Chapter 16.

Cloze procedure was used as a second, more demanding measure of language comprehension (see Klare, Sinaiko and Stolurow, 1971). Cloze procedure, which consists of systematically deleting every *n*th word (5th in our case), requires readers to fill in the blanks. The number of blanks for each passage-language combination differed for the following reasons : (1) slight differences in the lengths of the English passages, plus the accepted cloze measurement procedures of counting hyphenated words as separate words for deletion purposes, unless the compound contains a bound morpheme (e.g., «co-chairman»), and of excluding words in headings; (2) the fact that Vietnamese typically consists of several short, single-syllable words in place of a single, longer multi-syllable English word: and (3) the translation methods produced different numbers of words; e.g., ratios of words in human translation to computer translation were about 1 : 1.2. The number of blanks in each reading sample is summarized in the following table.

TABLE 1.

CLOZE BLANKS PER READING PASSAGE

| | Reading Passages | | |
|---|---|---|---|
| Versions | C1 | C3 | C16 |
| English (EN) | 107 | 103 | 105 |
| Human (HU) | 170 | 157 | 183 |
| Post-edited (PE) | 146 | 132 | 150 |
| Un-edited (UE) | 148 | 136 | 148 |

Two scores were derived from cloze tests: proportion of correct responses and proportion of omissions or blanks left unfilled. Proportions were used rather than raw scores to normalize the differences between passages in number of words. The common procedure in cloze measurement was followed: responses were counted as correct if they were misspelled; synonyms were not accepted.

b. *Rate of Work.* Time was recorded by the subjects themselves as they started and finished each of the separate activities they performed: reading the passages, taking the tests, and making the judgments. These made possible elapsed time measures for each activity.

c. *Judgments of Intelligibility (Clarity).* Subjects judged the clarity of each passage they read by assigning a rating from 9 («Perfectly clear and understandable. Sounds good to a reader.») to 1 («Not under standable at all. No amount of study would help a reader know what the main idea is». This «clarity scale» was based on the «scale of intelligibili ty» developed by John Carroll (1966) and rewritten for easier under standing by our VNAF subjects. (The original Carroll scale was intended for use by trained raters.)

3. *Experimental Design*

The 4× 3 (versions × passages) factorial design shown below was used for the study. There were twelve unique experimental conditions for the Vietnamese subjects (only the three English conditions applied in the case of USAF subjects).

For each of the cells shown in Table 2 we collected data on the several dependent measures described above. Since both the cloze test and the reading test could not be given to the same subjects on the same passage, it became necessary to give a subject one passage followed by the reading test, then material from a different passage in cloze format. Thus, our interest centered in 4×1 (versions × passage) univariate analyses for each dependent variable : reading lest score; cloze test scores (both percent correct and items omitted); time spent in reading a passage; time spent in taking the reading test; time spent in taking the cloze test; clarity scale score for a full-text passage; and clarity scale score for a passage in cloze format.

TABLE 2.

EXPERIMENTAL DESIGN

| | Reading Passages | | |
|---|---|---|---|
| Versions | Chapter 1<br>(C1) | Chapter 3<br>(C3) | Chapter 16<br>(C16) |
| English (EN) | × | × | × |
| Human (HU) | × | × | × |
| Post-Edited (PE) | × | × | × |
| Un-Edited (UE) | × | × | × |

## 4. *Administration*

Materials were assembled into booklets corresponding to the unique conditions of the experimental design. Testing was done in large groups, with booklets being assigned randomly within each class of student pilots. For the VNAF readers there were twelve sub-groups; USAF subjects read only English, corresponding to the three conditions in the top row of Table 2. Each man read a passage of full text, rated it, took a reading comprehension test, read a different passage in cloze format and filled in the blanks, and made a second rating; time notations were also made at beginning and end points of each subtest.

Subjects were given written instructions on the sheets at the beginning of each new activity (i.e., reading, taking a test, making a judgment), so that the men could work through the booklets continuously once they began. Conspicuous boxes were provided on the sheets at the beginning and end of each activity, so that time could be recorded. Instructions were given in English when the test material was in English; when the passage was in Vietnamese, instructions, test and rating scale were also in Vietnamese.

The administration was preceded by a brief explanation in English of the purpose of the experiment, and this was followed by similar comments in Vietnamese by a VNAF liaison officer. The subjects were told that our interest lay in testing the materials, not in testing them as individuals. Consequently, they were asked not to put their names on the booklets.

## III. RESULTS

### *Introduction*

The main emphasis in this chapter is the analysis of data collected on VNAF subjects. Comparisons with the control group of USAF subjects also appear in this chapter. The analysis is organized in terms of the eight dependent variables or measures. In all but three cases there were 14 subjects contributing to each condition. Data were lost for 9 subjects in the case of ratings of cloze passages, and for 16 subjects for time to complete cloze forms.

All the data have been subjected to rigorous statistical testing, including both multivariate and univariate analyses of variance, where appropriate. [4]) Thus, we refer to «significant» differences in the conventional or statistical use of the term; namely, that observations so labeled can be accepted with high confidence. Put another way, such measured differences can be considered to be highly reliable and not likely to occur as chance or random fluctuations more than 1 in 100 times.

### *Reading rate*

We feel that this is a critical variable in measuring the quality of written material, particularly when that material may be in a foreign language — English, in the case of VNAF readers—or when it is produced by any of several translation processes. Although our observations are based on three 500-word samples in each of the main language conditions, even

[4]) Significance testing followed Dunn's procedure (see Kirk, 1968).

slight but statistically significant differences have important imputations. An average difference of 1.5 minutes in reading two 500-word samples could result in many hours gained when summed over the tens of thousands of words that may be required reading in some military training programs, e.g., about 75 minutes for 100 pages of typed text.

Our observations showed significant differences in reading rate between certain of the language conditions, in particular, between the two fastest conditions (un-edited machine translation and human translation) versus the slowest (English). Table 3 summarizes average reading speed for each of the main conditions. (Differences between passages were not significant.) The slight mean differences between the three Vietnamese chapters were not significant.

TABLE 3.

READING SPEED : FOUR LANGUAGE CONDITIONS
(mean rate for three passages)

| Human Translation | Un-edited Machine Translation | Post-edited Machine Translation | English |
|---|---|---|---|
| 4.2 min | 4.1 min | 4.7 min | 5.7 min |

Figure 1 shows reading speed, both in terms of language conditions and the three chapters we sampled. Clearly, reading in a second language (i.e., English) results in a significant slowdown. This effect becomes most pronounced when VNAF subjects are dealing with the most technically complex material, i.e., Chapter 16. We do not have a ready hypothesis for explaining the slightly slower rates for the post-edited MT material.

*Reading comprehension*

Understanding was measured by fill-in tests for each chapter, administered in an open-book mode. Mean scores for the chapters ranged from 3 to 10.5. Overall rank order for the four language conditions was English (best performance), human translation, postedited MT, and un-edited MT (poorest). Differences were significant for both main variables, i.e., language condition and chapter. Figure 2 shows these results graphically.

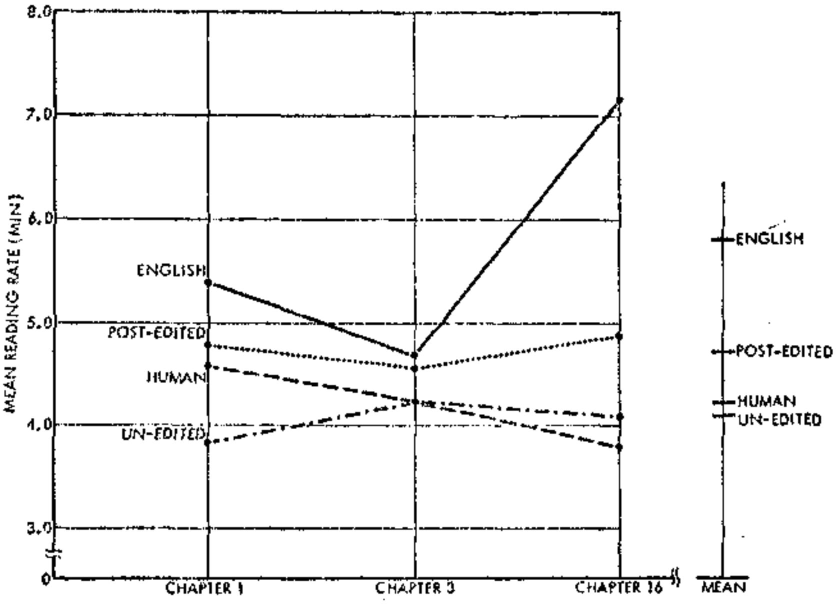

FIGURE 1. Reading Speed

Considering the language conditions, overall comprehension of English and of human-translated material differed slightly and *not* significantly; there was a tendency for the most complex material (Chapter 16) to result in higher scores when subjects read Vietnamese translated by manual means. Both English and human versions were significantly better than either MT version.

Most striking, as seen in Figure 2, is the very poor comprehension of un-edited MT for Chapter 16. Performance under all language conditions was surprisingly similar, and *not* significantly different, for the tests on Chapters 1 and 3. This suggests that some material might be left un-edited, particularly if it is not too technical.

*Judgments of clarity*

Subjects filled in 9-point rating scales to indicate their opinions of the intelligibility or clarity of each passage. Rank order, for the four language conditions, was: English (highest rating), post-edited MT, human

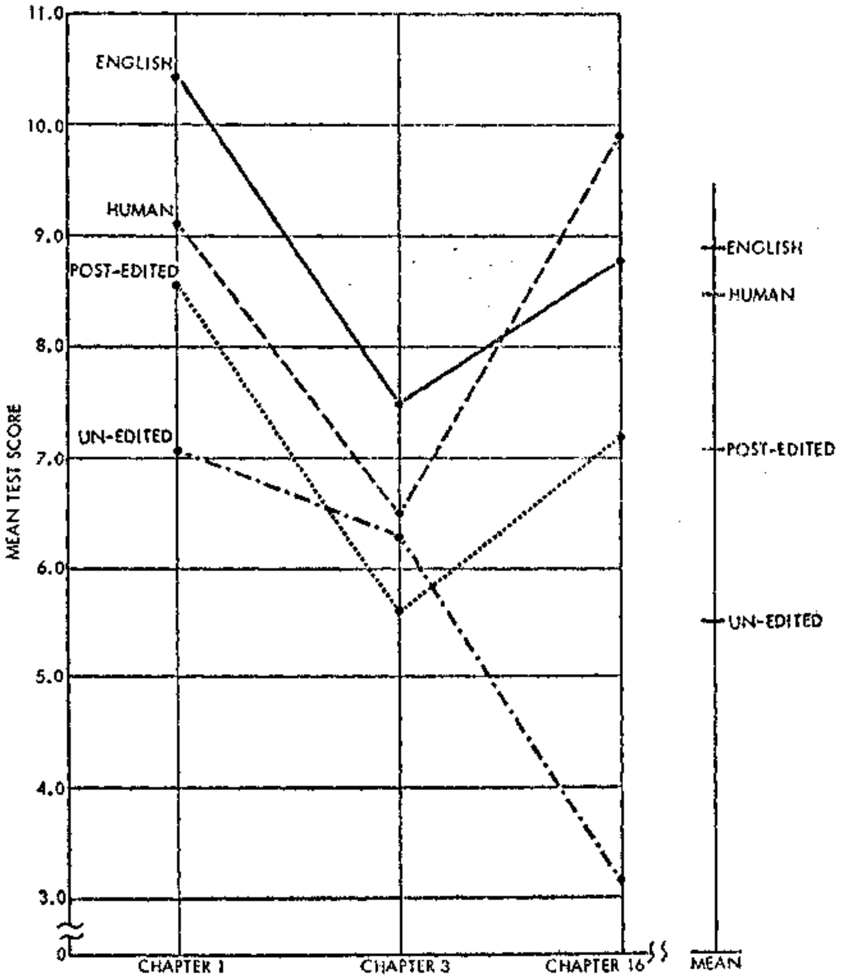

FIGURE 2. Reading Comprehension

translation, and un-edited MT. The only statistical significance lay in the differences between the lowest ranked and each of the two top-ranked versions. Figure 3 shows average clarity ratings by language and chapter variables. It is apparent from Figure 3 that un-edited MT was consistently perceived as least intelligible, and that ratings became successively lower with the more technically difficult material. It is notable that the human translators considered un-edited MT to be poor stylistically; however, post-edited MT was judged as being superior to the human translation by VNAF reader-subjects.

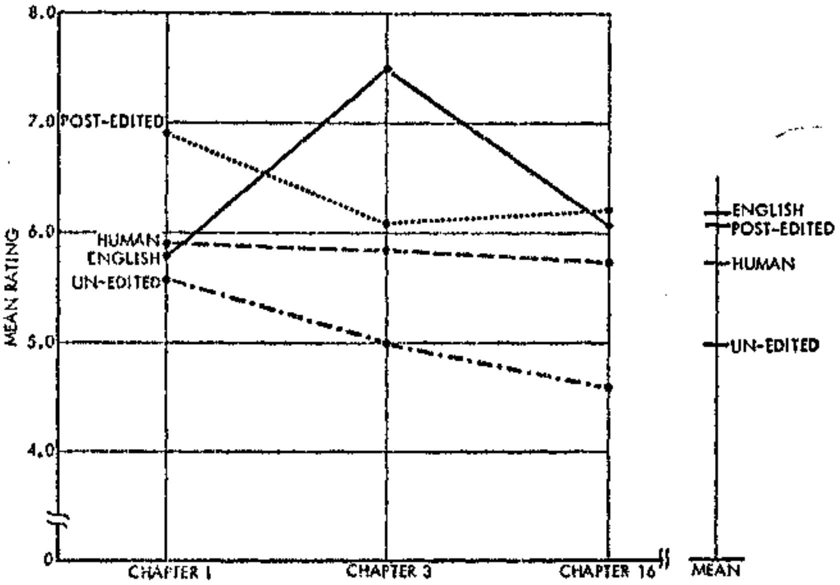

FIGURE 3. Clarity Ratings

*Cloze procedure scores*

We derived two measures of readability from the cloze tests: (1) correct responses and (2) omissions. Earlier experience (Klare et al., 1971) has shown that cloze is sensitive to different quality levels of translations of English-to-Vietnamese. The test is a rigorous measure of a reader's understanding of language because it. measures both «content» words (e.g., nouns, verbs) and «structure» words (e.g., articles, prepositions, etc.) (The practical significance of these in language use is discussed in Chapter IV.)

Figure 4 shows cloze scores for the four language conditions and for each of the three sampled chapters. Overall mean differences between languages (shown to the right of Figure 4) were all statistically significant. Also, average scores for Chapter 16 material (most technically difficult) were significantly lower than for either of the other passages. In addition, there were significant differences, *within* each of the passages, between languages. Finally, for the unedited and English material cloze mean accuracy scores varied significantly across chapters. (Note the two lowest curves in Figure 4.)

We believe that the relatively like performance (high) with human translation and post-edited MT indicates the essentially similar ability of those methods to deal with English-to-Vietnamese. It is also of interest that average scores tended to remain high as material became more complex, i.e., progressed from Chapter 1 to Chapter 16. On the other hand, cloze scores were relatively low for both English and un-edited MT. The latter, in particular, deteriorated between non-technical (Chapter 1) and technical material (Chapter 16). The striking similarity of English and un-edited MT scores on Chapter 16 suggests that our VNAF subjects, reading un-edited MT, produced the same low scores as they did for a foreign language, i.e., English. Low cloze performance throughout, using English text, is to be expected since this measure of readability

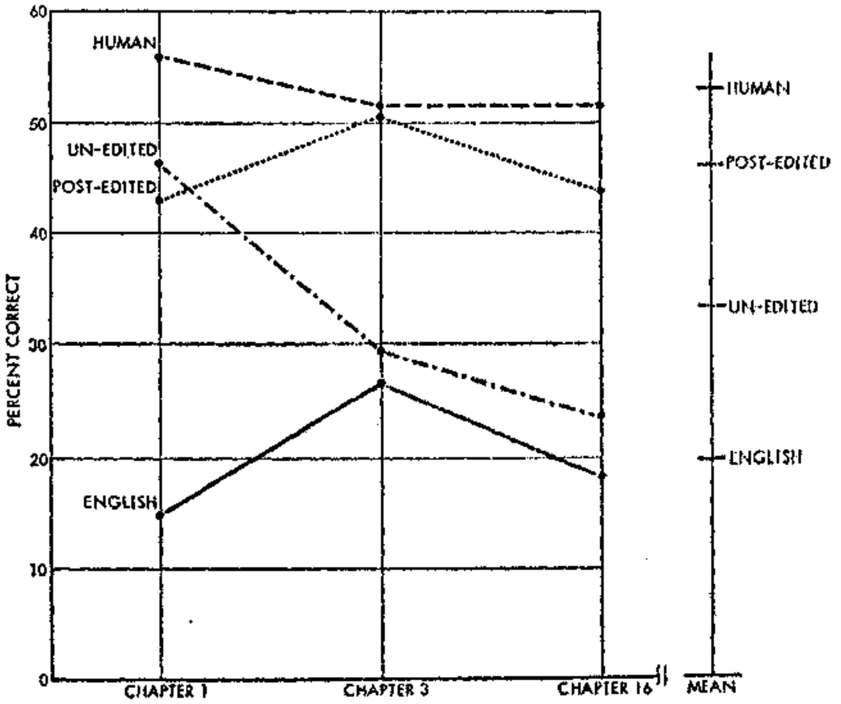

FIGURE 4. Cloze Procedure : Accuracy

is highly sensitive to idiom and other linguistic refinements likely to be difficult for readers in a second language.

The second measure derived from cloze tests was an index of difficulty determined by tallying unanswered or blank responses. These scores can be interpreted as indicative of a sort of ultimate difficulty in that respondents are unable to fill in any term at all in the cloze blank. Figure 5 shows the omission percentages for each of the experimental conditions.

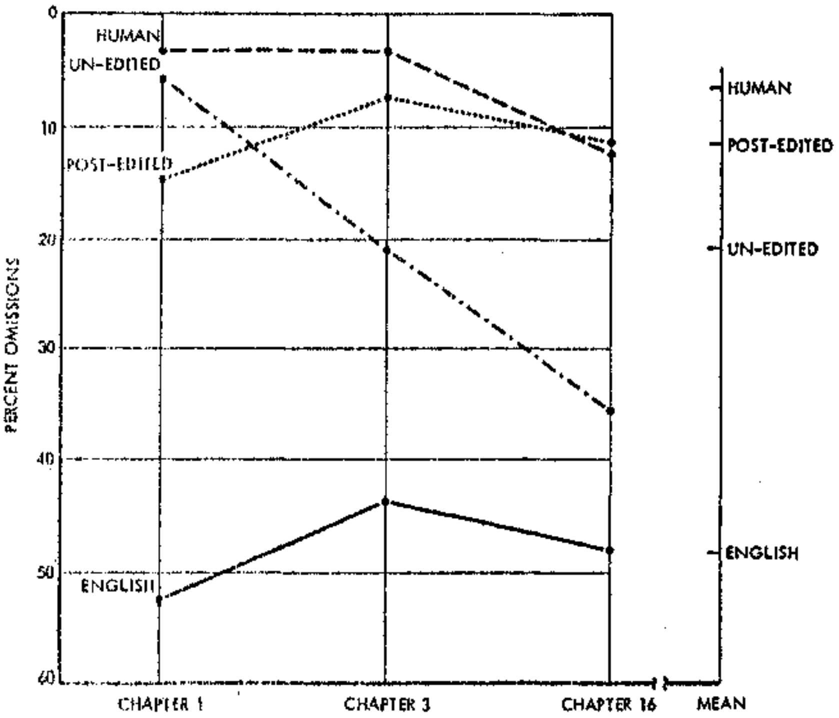

FIGURE 5. Cloze Procedure : Omissions

Significant differences occurred between English scores and each of the three translations. (There was borderline statistical significance between the un-edited MT scores and human translations.) As indicated earlier, one would expect the percentage of English word omissions to be high because of the nature of cloze procedure. Figure 5 also shows the relatively poorer performance with the more technical material when un-edited MT is used.

*Minor measures*

In addition to the above observations, we collected data on three additional aspects of readability: (1) time to complete reading comprehension tests, (2) time to complete cloze forms, and (3) clarity ratings of cloze material. Testing time, the first of these measures, did not discriminate among the four language conditions. That is, mean time to complete the tests for the three translations and the English test form differed only slightly from one another. Averages, in minutes, were: 10.4, 12.9, 11.4, and 11.2 (human translation, post-edited MT, un-edited MT, and English, respectively). We infer from these observations that searching for answers, as apart from comprehension, was not influenced by the language condition.

There were significant differences, however, between chapters. Reading tests for the *least* technical material (Chapter 1) were completed faster (mean time was 9.8 minutes) than for either of the other chapters (12.1 and 12.4 minutes, Chapters 3 and 16, respectively).

The second minor measure was time spent filling in cloze forms. Although English material tended to be handled most rapidly of the four language versions, differences were slight and not statistically significant. Recalling the high incidence of omissions with English cloze forms previously mentioned, the relatively short time spent filling them in is explained: readers simply did not understand many of the refinements of a foreign language, they left blanks, and their overall lest time was fast. Mean cloze time for the four language versions was: 32.1, 34.1, 36.5, and 30.5 minutes (human translation, post-edited MT, un-edited MT, and English, respectively). There were significant differences across passages; the tests based on material from Chapter 3 were completed more rapidly than those for Chapter 16,

Subjects rated the clarity or intelligibility of cloze formatted material using the same 9-point scale described above. Rank order, from most to least preferred versions, was: human translation, post-edited MT, English, and un-edited MT. The only statistical significance lay in the differences between each of the first two versions and the last-ranked. Figure 6 shows the clarity ratings of cloze material.

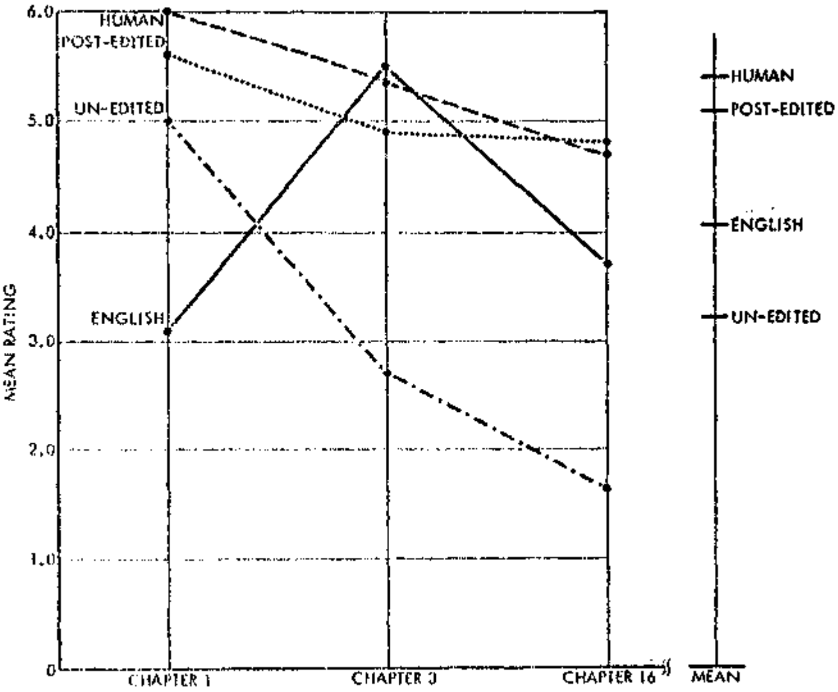

FIGURE 6. Clarity Ratings : Cloze Format

*Relationships among measures*

A large number of dependent measures (or test scores) were used in this study: reading test scores and cloze scores of two kinds; two kinds of ratings; and five kinds of time measures. Were they all necessary? Or were some measures so closely related to each other that not all were really needed? We ran a series of correlations to see. In general, we found that the correlations were low and inconsistent, suggesting that the various measures were independent (i.e., measuring different characteristics), and that all could therefore be usefully included. We must qualify this statement somewhat, however, in light of the number of cases (subjects) tested. We had 14 in each group, which is relatively small for a correlational analysis, and it is possible that relationships might have shown up more clearly if we had been able to test more persons. For example, the slight but consistent positive relationships found between clarity ratings on the full text and reading test scores might have been higher with larger groups

We did find one set of relationships that was consistently high despite the number of cases. It was that between number of cloze items correctly filled in and number of cloze items left blank, with correlations of -0.77, -0.98, and -0.95 on the groups having the English versions of Chapters 1, 3, and 16, respectively. These correlations are high enough, especially for technical Chapters 3 and 16, to suggest that, for a quick check at least, one might use a simpler «blank» count in place of a more time-consuming «correct» count. (Provided, of course, that the testing conditions and the subjects are similar in characteristics to those we had.)

### *English as a medium of instruction*

This section is a brief comparison of the performance of VNAF readers of English with USAF subjects. The latter group consisted of 88 student pilots who read the same material and were tested in the same way we measured readability among VNAF subjects.

The most rapid overall reading speed in the experiment was recorded for USAF subjects: 3.7 minutes mean time for three 500-word passages. Reading the same English material, the VNAF average was 5.9 minutes. (The fastest condition for Vietnamese text occurred with un-edited MT text: 3,9 minutes.)

Reading comprehension scores, in English for the three chapter samples, are shown in Table 4. Differences were slight and they did not consistently favor either group. Thus, VNAF subjects scored higher on the reading test for Chapter 1, while USAF subjects had higher scores on the remaining two passages.

TABLE 4.

READING COMPREHENSION :
ENGLISH VERSION, USAF AND VNAF SUBJECTS
(Mean Scores %)

| | Reading Passages | | |
|---|---|---|---|
| Group | C1 | C3 | C16 |
| USAF | 9.7 | 10.3 | 9.6 |
| VNAF | 10.5 | 7.5 | 8.8 |

Cloze scores, as shown earlier, were significantly much lower in the case of VNAF readers of English than for the American control group. Table 5 summarizes these data. We have indicated earlier that cloze procedure puts a major emphasis on understanding the structure of a language, hence the poor performance of VNAF subjects is not unexpected. This hypothesis is supported by an analysis of cloze scores comparing VNAF subjects with most and least exposure to English. (Beyond basic English instruction in Vietnam, as well as brief refresher training at Lackland AFB, our VNAF readers had been at Keesler from 3 to 27 weeks. This analysis is based on small groups of men at each extreme, i.e., 10 men who had more than 22 weeks of instruction and 10 with three weeks in the course.) Average cloze performance—percent of correct responses—was: 23.5 percent and 7.1 percent, respectively, for the men with maximum and minimum exposure to English.

TABLE 5.

CLOZE PROCEDURE :
ENGLISH VERSION, USAF AND VNAF SUBJECTS
(Mean Accuracy Scores %)

| | Reading Passages | | |
|---|---|---|---|
| Group | C1 | C3 | C16 |
| USAF | 57.3 | 58.1 | 52.1 |
| VNAF | 15.0 | 26.5 | 18.3 |

Clarity ratings were consistently higher when USAF readers evaluated the sample passages; overall means on the 9-point scale were 7.8 versus 6.5 for USAF and Vietnamese readers of English, respectively.

In summary, VNAF subjects read English material much more slowly than USAF controls but there was relatively little loss of reading comprehension compared to the Americans. (For one of the sample chapters, VNAF mean scores were higher than the USAF average.) More fundamental understanding of English structural aspects, as measured by cloze procedure, was significantly poorer among the VNAF subjects. There was some evidence that time-in-country was a major factor affecting cloze performance.

*Summary of results*

A relative comparison of VNAF scores on English and the three translation versions is shown in Table 6, with all other scores given relative to the highest score, which is labeled 100 %. We do not believe an overall or combined figure-of-merit would be meaningful in trying to assess the relative values of these modes of presenting training material, especially in view of the attempt to establish the human translation as a standard of excellence against which to compare the other conditions. Therefore, each measure should be considered separately. However, we would again call attention to the surprisingly high standing of English performance, with the suggestion that the use of this language not be summarily dismissed for subjects who have had some training in English.

TABLE 6.

RELATIVE PERFORMANCE OF VNAF SUBJECTS ON FOUR LANGUAGE VERSIONS (Percent)

| | Language Versions | | | |
|---|---|---|---|---|
| Measure | English | Human Translation | Un-edited Computer Translation | Post-Edited Computer Translation |
| Reading Comprehension (correct answers) | 100 | 96.5 | 59.7 | 79.8 |
| Cloze: Accuracy (correct answers) | 37.6 | 100 | 62.5 | 86.4 |
| Difficulty (omissions) | 55.2 | 100 | 84.0 | 94.0 |
| Reading Rate | 70.6 | 96.6 | 100 | 85.7 |
| Judgments of Clarity | 100 | 90.1 | 77.8 | 98.9 |

## IV. IMPLICATIONS

This chapter discusses the results of our experiments in terms of two things: (1) applications of our findings to training and translation problems and (2) methodological aspects of the study, and ways to improve some of them, that might be of help in other similar research.

### *Application to training and translation*

Translation by computer, or machine translation (MT), is surprisingly good from a research and development point of view. It is encouraging, we believe, that the present state of technology permits fairly sophisticated technical English to be processed by MT; resulting translations into Vietnamese can be read and understood by native readers of that language. However, when compared with excellent translations by human linguists the readability of MT output is inferior. In terms of dealing with very large volumes of English training and maintenance material (i.e., many thousands of pages and hundreds of thousands of words) the obvious advantages of MT are: (1) great speed in the central processor, (2) use of standardized lists of terms, and (3) a standardized output formal. The main disadvantages of MT, for volume production, are: (a) the costs of post-translation editing and recomposition (which are not known specifically but which appear to be excessive at present), and (b) the preparation of software and related lexical material for the computer.

In our three-way comparison of translation modes—human translators versus un-edited machine translation versus post-edited translation—readability measures consistently favored the first, with edited and un-edited material ranking second and third on most measures. Reading speed slightly favored the *un-edited* MT. We can only speculate why this was so; perhaps the un-edited material, which fared least well in measures of reading comprehension, was simply skimmed more rapidly. Informal subjective opinions of the translators indicated that the un-edited MT was very poor stylistically, i.e., it lacked the «flow and balance» that characterizes good literary Vietnamese. Nevertheless, our VNAF readers were able to reach a minimum level of comprehension of that material.

Translation readability varied somewhat depending upon the part of the English text on which the translation material was based. Higher readability scores tended to be associated with earlier chapters in the Air Force manual, particularly the first chapter which was the least technical of the three we sampled. Conversely, lowest readability scores occurred with the most technical material, particularly for the machine translation versions.

One unexpected result of the experiment was the performance of VNAF subjects who read and were tested in English. Average reading comprehension scores for these men were slightly higher than for the best (human) translation. When the comparison was made, controlling for differing amounts of exposure to English or time in the United States, the performance on English was even more striking; the men who had been here for five or six months did almost as well in English comprehension tests as the control group, i.e., USAF student pilots taking the tests in English. It is clear that proficiency in English, and consistent use of English in training, pays off significantly in terms of reading and understanding our training material. (This is consistent with remarks made by many USAF and U.S. Army officer-instructors with whom we have talked: the largest single problem in dealing with Vietnamese students is their lack of ability to handle English, particularly conversational English and written technical material.)

It has been pointed out that practical considerations limit the possibility of training 100 percent of the Vietnamese Air Force to read English. However, consideration should be given to providing this qualification to initial training cadres in the VNAF so that those men can serve to clarify questions of ambiguous or faulty translations. Such an approach assumes, of course, that the bilingual instructors have available both the original English texts and translations.

It was significant, we believe, that the different translation methods produced Vietnamese passages of widely different lengths: a 535-word English sample was translated manually into approximately 850 Vietnamese words, and 740 and 730 words, respectively, by MT unedited and post-edited. Recalling that the highest reading comprehension scores were made using the longest (manual) translation, we hypothesize that this might reflect a basic difference between human and computer translations: the longer translation (manual) contained greater redundancy, hence more information that resulted in higher comprehension scores.

### *Methodological issues*

As one of our measures of readability we used a so-called «clarity» rating scale. This was a 9-point scale that had been modified from original use in other research and that had been translated into Vietnamese for use with the translations. Despite our preparation of an instructional paragraph, and our verbal explanations both in English and Vietnamese, it became apparent that many of the subjects simply did not understand the rating scale method. This came out in several ways; some subjects did not check any of the rating points, some checked several, and some gave ratings that were negatively correlated with performance on the comprehension tests (e.g., subject gives a very high clarity rating to a passage on which he scored zero on a reading test). We submit that lack of familiarity with this type of verbal material was the cause of the difficulty; that is, the Vietnamese subjects had rarely, if ever, used rating scales and our explanations were not adequate. Unfortunately, we did not have the opportunity to pre-test the scale and its instructions.

Our use of the cloze procedure, as one of the two main readability measurement techniques, was unduly harsh for Vietnamese readers of English. That is, cloze responses in a second language had the lowest accuracy scores of the live main language variables: 19 percent versus 56 percent for Americans working in English. We think that the reason for the poor performance, using cloze as the dependent measure, is a function of the technique. The cloze procedure deletes many «content» words (i.e., nouns, verbs, adjectives, adverbs), which non-native readers tend to be taught and/or learn first when [hey study a new language. They can, therefore, correctly fill in these words relatively well, which is also shown by the high scores on our fill-in reading test, which tended to use such words. However, cloze also deletes many «structure» words (i.e., articles, conjunctions, prepositions, exclamations, etc.), which play a large part in the idiom of the language. Non-native readers learn to use these words correctly only rather slowly, and therefore have relatively much more trouble filling them in properly than do native readers. The cloze technique thus appears to be a considerably more difficult task for non-native than for native readers, an observation not previously noted in the cloze literature.

All of our tests were taken in the language of the material being assessed. That is, reading test items were translated into Vietnamese (for those subjects who read translations) as were the rating scales. An obvious disadvantage (although worth mentioning and emphasizing) is that written responses had to be scored by native readers of Vietnamese. In retrospect, we suggest that items using the multiple-choice format be used instead of the fill-in style to obviate the need for costly test-scoring aides.

We used the subjects themselves to time the various parts of the test battery ; when the men came to a conspicuously labeled bow marked «TIME», they read their wrist watches and noted the hour and minute in the box. In general the technique worked satisfactorily but we did lose some data because subjects simply forgot to write the time (4 percent of our time scores). Added redundancy, in the form of time-measuring points spaced more frequently throughout the material, would have cut down on our lost data. For example, time boxes at both the end of one phase of activity and the beginning of the next would have provided for some forgetfulness.

## REFERENCES

Anonymous, *An Evaluation of Machine-Aided Translation Activities at FTD,* Arthur D. Little, Inc., Case 66556, May 1965.

BYRNE, C.E., B.E. SCOTT, and T.N. BINII, «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.

BYRNE, C.E., B.E. SCOTT, and T.N. BINII, «Optimization of LOGOS I System (Phase I),» Rome Air Development Center, Griffiss Air Force Base, New York, RADC-TR-70-270, December 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.

DINEEN, G.P., et al., «Final Report of the USAF Scientific Advisory Board Ad Hoc Committee on Mechanical Translation of Languages,»» Scientific Advisory Board, U.S. Air Force, September 1965.

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

HUTCHINSON, J.C., Defense Language Institute, Personal Communication, July 1971.

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 : Iowa State University Press, 1964.

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,

KROLLMAN, F., et al., «Production of Text-Related Technical Glossaries by Digital Computer. A Procedure to Provide an Automatic Translation Aid.» (Unpublished paper, translated from German; 1965).

LOGOS Development Corporation. Final Report on Air Force Contract No. F 30702-7l-C-0063. Middletown, N.Y., May 1971.

PIERCE, J.R., et al., L*anguage and Machines: Computers in Translation and Linguistics.* Publication 1416. Washington, D.C. : National Academy of Sciences — National Research Council, 1966,

SINAIKO, H.W., «Foreign Language Training: An Investigation of Research and Development for Vietnam, » Study S-232, Institute for Defense Analyses, Arlington, Va., March 1966.

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

United States Air Force, *Instrument Flying,* AF Manual 51-37, Washington, D.C. : U.S. Government Printing Office, 1968.