[*Mechanical Translation and Computational Linguistics*, vol.10, nos.1 and 2, March and June 1967]

Comprehensibility of Machine-aided Translations of Russian Scientific Documents*

by David B. Orr and Victor H. Small† American Institutes for Research, Washington, D. C.

This study used special reading-comprehension tests to compare the speed and accuracy with which the same Russian technical articles in physics, earth sciences, and electrical engineering could be read by technically sophisticated readers when they were presented in English translated from the original Russian by machine only, by machine plus post-editing, and by normal manual procedures. Thus, the emphasis was on the transmission of the technical message rather than on linguistic characteristics. In general, the results consistently showed that manual translations exceeded post-edited translations, which exceeded machine translations across all three disciplines and various types of questions. Losses in speed and efficiency were substantially greater than in accuracy, and differences between machine alone and post-edited generally exceeded differences between post-edited and manual translations. However, it was concluded that machine-alone translations were surprisingly good and well worth further consideration under the proper circumstances.

Problem

In the last one and one-half decades, there has been a growing interest in the use of computer-based techniques for the translation of foreign languages into English, particularly with respect to scientific and technical documents. During this period, rather large sums of money have been spent in the development and implementation of computer techniques for this purpose, while relatively little effort has been devoted to the evaluation of the outcome, at least from the point of view of communication of the technical material.

Reference to the literature of machine-translation research (see e.g., Edmundson[1] and See[2]) shows that virtually all of the research in this field, at least through 1964, has been concerned with the problems of developing computer configurations, dictionaries, syntactic and transformational processing, semantics, and similar hardware, software, or linguistic concerns. This work has obviously been essential to the development of machine translations against criteria derived from these disciplines to the neglect of evaluations based on the functional criteria of usability and comprehensibility. More recently, some research concerning the practice of machine translations has begun to appear (e.g., Pfafflin[3] and Carroll[4]).

The study reported here was of the latter type. Its principal objective was to compare by means of special reading-comprehension tests the accuracy and speed with which the same Russian technical articles could be read by technically sophisticated readers when they had been translated into English by means of two computer-based techniques and by normal manual translations. Thus, this approach differed sharply with most previous research in this area in that it placed primary emphasis on whether or not the technical message gets through in the translation process rather than on reactions to linguistic inelegance and linguistic inaccuracy.

Procedures

The study dealt with the comprehension of complete journal articles drawn from three technical fields: physics, earth sciences, and electrical engineering. A sample set of thirteen, eleven, and thirteen articles, respectively, was selected to provide a total of about twenty thousand words for each field. The articles were selected in collaboration with consultants to cover a range of significant topics within the field, to be primarily text rather than figures or tables, and to be as typical as possible of Russian journal content in that field.

An effort was made to use only articles which had been translated under the auspices of an American professional society. Each translation was checked and corrected by an independent, Russian-reading subject-matter consultant, to insure the best possible hand translation. Machine translations were produced by the Foreign Technical Documents Center of the Air Force at Wright-Patterson Field, Ohio, and represented the then current capability of that facility, which employed

* This work was performed in part under the sponsorship of the Air Force's Rome Air Development Center, Griffiss Air Force Base, New York, Contract No. AF30(602)3459. Copies of the full report may be requested from the Office of Information, Griffiss AFB. The assistance of the contract monitor, Mr. John McNamara, is gratefully acknowledged.

† Now with the Research Division, Montgomery County Schools, Maryland.

the IBM Mark II translation system.[5] Post-edited machine translations were used as the third translation condition, with the post-editing also being done by the FTD Center at Wright-Patterson. (An extensive analysis of FTD operations has recently been released by A. D. Little, Inc., 1966.[6]) Hand translations were either retyped or photographed for reproduction; post-edited translations were retyped; and machine translations were reproduced from the machine output. In the latter two cases, it was necessary to strip in graphs and figures from the originals.

The hand translations were used as the basis for test construction. Four-choice multiple choice items based on text rather than figural or pictorial material were written by a member of the staff expert in writing reading-comprehension tests. All sets of items were submitted to subject-area experts for- technical review. These items were designed to assess the general comprehensibility of articles. Some items were written to assess the transmission of factual material clearly stated in the text; some items paraphrased material stated in the text; and some items required the reader to draw inferences or interpret textual material.

About one item per hundred words of text was required for adequate coverage of the articles. In order to allow for refinement of the tests, the tryout forms contained 495, 549, and 445 items, respectively, for physics, earth sciences, and electrical engineering. Because of the length of these forms, the test material was divided into subtests which were counterbalanced in the pretesting to offset the results of fatigue and to permit some examination of results as a function of testing time. Answers to the questions were recorded in separate answer booklets.

The use of complete articles rather than selected passages (the usual procedure) required an additional innovation in test procedure. Pages of questions were interleafed with the pages of text from which they were drawn, and questions were keyed by numbers to the relevant paragraphs of text. Thus, in referring back to the text, the subject could avoid the extremely long and time-consuming search that would be necessary if all questions followed the article. It was felt that this innovation was essential not only for efficiency of testing, but also to maintain the motivation and interest of the subjects.

As an illustration of materials used in the study, a typical sample of text from the physics material is shown below in all three versions (machine, post-edited, and hand) along with the relevant questions.

SAMPLE OF MACHINE TRANSLATION

[§9]

Distinction ( ). Distinction in diffraction patterns, obtained at/during scattering of x-rays in layers isotope-in hydrogen, condensed on lateral surface of cold cylinder, it is possible uncontradictorily to explain by presence in such layers of texture and besides different for protium and deuterium. This isotopic effect in character of texture it is possible to compare with/from known from literature[3] temperature dependency of character of texture for is shell hexagonal metals, precipitated/deposited from vapor phase. Thus, for instance, zinc and cadmium at a temperature of sublayer higher than ~0.7t_M (t_M—melting point of corresponding metal) are crystallized with predominant orientation of plane (002) perpendicularly to sublayer (as also protium at/during 4.2° K), and at a temperature of sublayer lower 0.7t_M—with predominant orientation of this plane to in parallels to sublayer (how/as deuterium at/during 4.2° K).

[§10]

For protium and deuterium having different melting points and sharply different equilibrium vapor pressure at/during given temperature, sublayer with temperature 4.2° K possesses different effective temperatures. She/it effectively colder for deuterium than for protium. It is possible that namely this temperature dependency of texture one should explain isotopic effect in character of texture isotope-in-hydrogen.

SAMPLE OF POST-EDITED TRANSLATION

[§9]

The distinction in diffraction patterns obtained during scattering of X-rays in layers of hydrogen isotopes condensed on the lateral surface of a cold cylinder can be uncontradictorily explained by the presence in such layers of a texture different from protium and deuterium. This isotopic effect in the character of the texture can be compared with the temperature dependence known from literature[3] of the character of texture for layers of hexagonal metals, settled from the vapor phase. Thus, for instance, zinc and cadmium at a temperature of backing high than ~0.7t_M (t_M is melting point of corresponding metal) are crystallized with predominant orientation of plane (002) perpendicular to backing (as also protium at 4.2° K), and at a temperature of backing lower than 0.7t_M—with predominant orientation of this plane parallel to backing (as deuterium at 4.2° K).

[§10]

For protium and deuterium, having different melting points and sharply different equilibrium vapor pressure at a given temperature, a backing with a temperature of 4.2° K possesses different effective temperatures.

It is effectively colder for deuterium than for protium. It is possible that namely this temperature dependence of texture should explain isotopic effect in the character of texture of hydrogen isotopes.

SAMPLE OF HAND TRANSLATION

[§9]

The difference in the diffraction patterns obtained when x rays are scattered from layers of the hydrogen isotopes condensed on the side surface of a cold cylinder can be explained consistently by the presence of texture in such layers and by its difference for protium and deuterium. This isotope effect in the type of texture can be compared with the temperature variation, well known in the literature,[3]

in the type of texture in layers of the hexagonal metals deposited from the vapor phase. Thus, for example, at a substrate temperature above ~$0.7t_M$ (t_M is the melting temperature of the corresponding metal), zinc and cadmium crystallize with a preferential orientation of the (002) plane perpendicular to the substrate (as in protium at 4.2° K), and for a substrate temperature below $0.7t_M$ they crystallize with a preferential orientation of this plane parallel to the substrate (as for deuterium at 4.2° K).

[§10]

For protium and deuterium, which have different melting temperatures and sharply differing equilibrium vapor pressures at a given temperature, a substrate at a temperature of 4.2° K has different effective temperatures. It is effectively colder for deuterium than for protium. It is possible that the isotope effect in the texture type for the hydrogen isotopes should, in fact, be explained by this temperature variation of texture.

SAMPLE TEST QUESTIONS

[§9]

Zinc and cadmium resemble the hydrogen isotopes in having

A. a constant preferential orientation.
B. the same effective temperature.
C. isotopic polymorphism.
D. hexagonal crystals.

Which one of the following crystallizes with a preferential orientation of the (002) plane perpendicular to the substrate?

A. Zinc below $0.7T_M$
B. Zinc above $0.7T_M$
C. Cadmium below $0.7T_M$
D. Deuterium at 4.2° K.

[§10]

Variation in effective temperature may have led protium and deuterium to show different

A. atomic weight.
B. preferential orientation.
C. reactions to impurities.
D. numbers of sides in their lattices.

When protium and deuterium are condensed on the side surface of a cold cylinder, they may have different diffraction patterns because they have different

A. substrate effective temperatures.
B. substrate temperatures.
C. numbers of angles in their lattices.
D. degrees of chemical reactivity.

The tryout forms were administered as power tests essentially untimed) to fifty, forty-five, and thirty-five graduate students in physics, earth sciences, and electrical engineering, respectively. These students were paid twenty-five dollars for the testing which took four to eight hours. The typical item statistics were computed for these pretest data: item difficulties, Kuder-Richardson reliabilities, and item-test correlations. These statistics were used to select the items for the final forms of the test. Items were retained in such a way as to maintain coverage of the text. Those items passed by virtually all subjects, and those showing a negative correlation with total test score were eliminated.

The final forms of the tests were also subjected to item analyses. The characteristics of the tests are shown in Table 1. It can be seen that the tests tended

TABLE 1

ITEM STATISTICS, FINAL TEST FORMS

FIELD	*N* ITEMS	r_{xx}*	TRANSLATION TYPE Hand	Post-edited	Machine
Physics	221	.92			
Median difficulty			.88	.82	.75
Median item-test *r*†			.57	.57	.58
Earth sciences	189	.92			
Median difficulty			.86	.85	.76
Median item-test *r*†			.56	.47	.57
Electrical engineering	225	.91			
Median difficulty			.65	.60	.50
Median item-test *r*‡			.32	.33	.29

* Kuder-Richardson (No. 20) subtest reliabilities corrected to full length tests by the Spearman-Brown Formula,
† Biserials computed against article total scores.
‡ Biserials computed against subtest total scores.

to be somewhat easy. This was a deliberate device to maintain motivation. (However, the electrical engineering test was made somewhat more difficult by a decision to use more items requiring inference, as compared to direct factual or paraphrased items.) Final distributions had sufficient variance for analysis. The K-R reliabilities were based on subtests formed for purposes of the design (see below). When corrected to full length, they were deemed quite satisfactory.

In addition to supplying the necessary item statistics to construct the final test forms, the pretest data also provided information about test performance as a function of testing time. In general, these analyses indicated that subjects increased their working speed significantly while comprehension accuracy declined slightly over time. Accuracy rate scores generally improved with practice. These changes were modest, of the order of 1-2 per cent. There were differences in performance as a function of half-tests, however, indicating that half-test content and/or characteristics of the comprehension-test questions may have influenced performance scores. The fact that no serious losses in performance occurred as a function of time speaks extremely well for the level of motivation of these subjects, many of whom spent almost a full working day taking their respective tests. This observation lends considerable weight to the stability of the findings of the study in general.

TABLE 2
EXPERIMENTAL DESIGN

	PHYSICS (N=120)			EARTH SCIENCES (N=144)			ELECTRICAL ENGINEERING (N=120)		
	Subtest			Subtest			Subtest		
BOOK	1	2	3	1	2	3	1	2	3
Article numbers	1-4	5-8	9-13	1-4	5-7	8-11	1-5	6-9	10-13
1	Hand	Post-ed.	Machine	Hand	Machine	Post-ed.	Hand	Post-ed.	Machine
2	Machine	Hand	Post-ed.	Post-ed.	Hand	Machine	Machine	Hand	Post-ed.
3	Post-ed.	Machine	Hand	Machine	Post-ed.	Hand	Post-ed.	Machine	Hand

Experimental Design

For each discipline, the total test was subdivided into three parts, or subtests of as nearly equal length as the variety of article lengths permitted. Three different subtest books were constituted by assigning the three translation types of each subtest in a differing arrangement. Each book contained a subtest with hand-, post-edited, and machine-translated tests.

The set of three test books thus provided a partially counterbalanced, Latin Square arrangement in which each translation type was used in the early, middle, and late test period, as a control for learning and fatigue effects. Since these effects were counterbalanced across the three different groups of test subjects, it was necessary that the subject groups be constituted so as not to differ significantly in background and ability. Test books were assigned to subjects at random so that there was no known systematic bias upon which test groups could be distinguished. The design is summarized in Table 2.

For the final testing, only volunteers, advanced graduate students in the appropriate fields, were employed. Testing arrangements were made through university department heads and testing was carried out at about thirty universities across the country. Subjects were paid twenty dollars to twenty-five dollars for their participation. Testing sessions were held either on subsequent Saturdays or, for electrical engineering, all on a single day. Subjects were instructed to work at a good speed and to attempt each question in turn, but not to spend an unreasonable amount of time on any one question. All items were to be answered, even if guessing was required. The subject was asked to circle the number of the item upon which he was working at the sounding of a bell or buzzer at the end of each 10-minute interval. Mid-morning or mid-afternoon break periods were provided.

Each test was set up to obtain three scores. Since

TABLE 3
UNADJUSTED PHYSICS MEANS AND STANDARD DEVIATIONS FOR THREE TRANSLATION TYPES (N = 120)

	TRANSLATION TYPE						
	Hand		Post-Edited		Machine		MEAN
SCORE AND SUBTEST	Mean	s	Mean	s	Mean	s	TOTAL
% Correct by subtest:							
1	84.69	7.60	80.51	9.31	75.03	12.24	80.08
2	83.38	7.25	85.04	6.22	78.91	9.40	82.44
3	82.60	8.20	77.34	9.50	72.86	9.91	77.60
Total	83.56	7.68	80.96	8.99	75.60	10.80	80.04
N 10-min. intervals by subtest:							
1	9.70	2.17	11.72	2.94	11.72	3.31	11.05
2	7.22	1.25	8.42	1.63	10.67	2.93	8.77
3	9.05	1.92	9.10	1.84	10.67	2.08	9.61
Total	8.66	2.09	9.75	2.62	11.02	2.34	9.81
N correct/10-min. interval by subtest:							
1	6.73	1.77	5.36	1.60	4.96	1.35	5.68
2	8.41	1.44	7.44	1.57	5.61	1.57	7.15
3	7.26	1.73	6.65	1.29	5.32	0.99	6.41
Total	7.47	1.78	6.49	1.71	5.30	1.34	6.42

TABLE 4
UNADJUSTED EARTH SCIENCE MEANS AND STANDARD DEVIATIONS FOR THREE TRANSLATION TYPES (N = 144)

SCORE AND SUBTEST	TRANSLATION TYPE						MEAN TOTAL
	Hand		Post-Edited		Machine		
	Mean	*s*	Mean	*s*	Mean	*s*	
% Correct by subtest:							
1	78.09	11.52	73.57	9.54	69.04	10.30	73.57
2	82.09	9.24	82.39	7.40	68.85	11.08	77.78
3	78.41	8.57	71.33	8.87	63.36	10.70	71.03
Total	79.53	9.96	75.76	9.84	67.08	10.95	74.13
N 10-min. intervals by subtest:							
1	7.50	2.03	8.71	2.16	9.65	2.86	8.62
2	7.23	1.59	7.35	1.41	8.25	2.09	7.61
3	7.00	1.29	8.46	1.62	9.54	2.02	8.33
Total	7.24	1.67	8.17	1.84	9.15	2.43	8.19
N correct/10-min. interval by subtest:							
1	7.10	1.98	5.70	1.43	5.01	1.73	5.94
2	7.32	1.57	7.17	1.39	5.43	1.36	6.64
3	7.32	1.68	5.52	1.33	4.31	0.93	5.72
Total	7.25	1.74	6.13	1.56	4.91	1.45	6.10

the test was a power test, an accuracy score, or a measure of extent of comprehension of the material, was defined as the percentage of correct answers to the total number of questions asked. The second score which was obtained was the total amount of time taken to answer the items in the test in terms of the total number of 10-minute periods taken to answer the test items. The third measure, accuracy rate, was defined as the number of items correct per 10-minute period. This score represented an efficiency statistic indicating the extent to which the type of translation could be used to get correct information in a comparatively short time.

Results

The analysis of variance approach was used to determine whether there were statistically significant differences attributable to the variable of interest. The same

TABLE 5
UNADJUSTED ELECTRICAL ENGINEERING MEANS AND STANDARD DEVIATIONS FOR THREE TRANSLATION TYPES (*N* = 120)

SCORE AND SUBTEST	TRANSLATION TYPE						MEAN TOTAL
	Hand		Post-Edited		Machine		
	Mean	*s*	Mean	*s*	Mean	*s*	
% Correct by subtest:							
1	63.63	7.91	58.20	8.47	54.47	6.80	58.77
2	65.17	9.81	63.90	11.70	51.03	10.98	60.03
3	60.07	11.74	59.80	9.24	51.00	10.10	56.96
Total	62.96	10.09	60.63	10.11	52.17	9.53	58.59
N 10-min. intervals by subtest:							
1	12.30	3.12	13.00	3.23	14.63	3.97	13.31
2	10.90	2.07	11.50	2.41	12.02	2.87	11.47
3	9.17	1.96	9.17	1.74	10.55	2.46	9.63
Total	10.79	2.74	11.22	2.97	12.40	3.56	11.47
N correct/10-min. interval by subtest:							
1	4.11	1.10	3.54	0.95	2.97	0.80	3.54
2	4.60	0.94	4.32	1.10	3.30	0.84	4.07
3	5.09	1.32	5.03	1.09	3.79	1.03	4.64
Total	4.60	1.19	4.30	1.20	3.36	0.95	4.08

basic Latin Square design was used throughout.[7] Where the analyses indicated that a significant effect attributable to type of translation did exist, Duncan tests[8] were performed to determine where these differences lay. (The Duncan test is a modified t-test for testing the significance of differences between three or more means to show whether every mean is different from every other mean or whether there are significant differences between some means and not between others.)

Direct comparisons of subject fields should not be made since the numbers of items in the tests differed and since the tests were not equated in difficulty or content.

Means and standard deviations for the basic data are shown in Tables 3, 4, and 5. Analyses of variance were carried out to test the differences in translation types for each discipline. These analyses are summarized in Table 6.

COMPREHENSION ACCURACY

The accuracy trends for subtests within disciplines and for the three disciplines were markedly similar. Simple differences in percentage accuracy between hand and post-edited translations consistently ranged from 2.6 per cent to 3.8 per cent across all analyses, significant statistically except for electrical engineering. Differences between post-edited and machine translations were also consistent, significant, and somewhat larger. The range of simple differences in percentage accuracy across all analyses was from 5.4 per cent to 8.7 per cent for post-edited versus machine translations. The differences in accuracy between hand and machine translations were both consistent in direction and more substantial in magnitude and were significant statistically. They ranged from 8.0 per cent to 12.5 per cent.

RATE OF WORK

All translation comparisons among mean time scores were significant for physics and earth sciences. For electrical engineering, the time required for hand versus post-edited translations did not achieve significance. The difference between hand and machine translation times ranged from 24.0 to 16.1 minutes per subtest across all disciplines.

ACCURACY RATE

For all groups tested, the differences between the means for hand and machine and between post-edited and machine translations were consistently significant and ranged from 1.2 to 2.2 items correct per 10-minute period. The differences between hand and post-edited translation means were not significant for electrical engineering.

RELATIVE LOSSES WITH POST-EDITED AND MACHINE TRANSLATIONS

The analyses reported above indicate the direction, extent, and statistical significance of the differences between mean criterion measures for the three translation types being compared. In addition, the *relative* differences in mean scores between hand translations and both post-edited translations and machine translations were computed for all test groups. (Percent difference = 100—[X comparison/X standard] 100 where scores are directly related to efficiency and 100 $[X_c/X_s]$—100 where scores are inversely related to efficiency.) They indicate percentage losses in accuracy, percentage increases in time required per item, and percentage reduction in the number of items correct per unit of time *where the hand translation was*

TABLE 6
SUMMARY OF ANALYSES OF VARIANCE BY SCORE AND DISCIPLINE

	PHYSICS				EARTH SCIENCES				ELECTRICAL ENGINEERING			
		F				*F*				*F*		
SOURCE	d.f.	% Correct	*N*	*N*/10 min.	d.f.	% Correct	*N*	*N*/10 *mm.*	d.f.	% Correct	*N*	*N*/10 min.
Between subjects:												
Groups	2	1.05	3..94*	2.31	2	3.10*	1.45	4.11*	2	2.09	...	...
Subjects within groups	117				141				117			
Within subjects:												
Type of translation	2	69†	62†	157†	2	169†	60†	187†	2	91†	162†	75†
Subtests	2	25†	59†	72†	2	48†	18†	32†	2	6.78†	79†	54†
Translation X subtest	2	4.17*	. . .	1.88	2	. . .	3.64*	2.35	2	1.63	2.10	1.56
Error (within)	234				282				234			
Total	359				431				359			

* Significant at the 5% level
† Significant at the 1% level.

used as a standard of comparison. All differences represent decrements of performance in relation to the standard. These relative performance losses for all disciplines are shown in Table 7.

It can be seen from Table 7 that the percentage loss in performance level for machine translations as compared to hand translations was two to three times as great for all three measures as the percentage loss for post-edited translations compared to hand translations. Furthermore, the greatest losses occurred in the measures of time required and number correct per unit of time, rather than in accuracy (per cent correct).

QUESTION-CATEGORY ANALYSES

In view of the variety of questions contained in the tests, it was of interest to make translation comparisons based on more homogeneous, more functional types of questions. The categories of questions used in these analyses were: (1) Literal-Direct: Statements or questions based on material presented directly and in full in the text; (2) Equivalent-Direct: Statements or questions covered in full in the text, but paraphrased or equivalently stated; (3) Indirect Inferential-Understanding: Statements or questions not covered directly in the text, but requiring the reader to comprehend the meaning of the material beyond a single word or sentence in order to infer, generalize, or integrate the materials contained in the text to produce the answer.

The question-category data are reported in terms of accuracy scores only, since the various categories of items were imbedded unsystematically in the total test, and no meaningful time measures could be obtained. The number of items in the three categories, respectively, for physics was seventy-four, seventy-three, and fifty-eight; for earth sciences thirty-five, ninety-one, and forty-two; and for electrical engineering thirty-seven, ninety-five, and ninety.

The results of these analyses are summarized in Figure 1. Subtest mean scores were adjusted to eliminate the group differences for plotting profiles of subtest means for each translation type, so that the plots represented the within-person subtest X translation interaction pattern as treated in the analyses of variance. Analyses of variance similar to those reported for the main analyses were also run, but are not shown here to conserve space. For all disciplines, the mean trend of accuracy scores showed overall a remarkable similarity to the findings of the main analyses. There tended to be a decline from hand to post-edited translations and a sharper decline for machine translations. For questions categories 1 and 2, three of the six comparisons were significantly different for hand versus post-edited translations. The trend, while similar for category 3 questions, was less marked; the differences were not significant. Accuracy for hand versus machine translations differed markedly for question categories 1 and 2 and differed almost as much for question category 3.

For all disciplines, there was a progressive reduction in accuracy from question category number 1 to 2 to 3. Thus, comprehension accuracy for questions involving paraphrased statements was lower than for questions involving direct statements and lower still for statements which required the subject to show understanding and/or to draw inferences based upon the textual material.

Most scientific articles can be divided into several sections of content. As a check on the item-category results above, items were reclassified into those dealing with the following sections of the articles: Problem, Background, Approach/Method, Results, Discussion, and Conclusions. The trend lines of these translation comparisons were found to be essentially similar to those described above. However, in these analyses, differences between hand and post-edited translations were less pronounced than before and sometimes in the opposite direction.

TABLE 7
PERCENTAGE DECREMENT IN CRITERION SCORES FOR POST-EDITED AND MACHINE TRANSLATIONS COMPARED TO HAND TRANSLATIONS AS A STANDARD FOR THREE DISCIPLINES

Score	Discipline	Post-Edited/Hand	Machine/Hand
Percentage correct	Physics	3.1	9.5
	Earth sciences	4.7	15.7
	Electrical engineering	3.7	17.1
N 10-min. intervals	Physics	12.6	27.3
	Earth sciences	12.9	26.4
	Electrical engineering	4.0	14.9
N corr./10-min. interval	Physics	13.1	29.0
	Earth sciences	15.4	32.3
	Electrical engineering	6.5	27.0

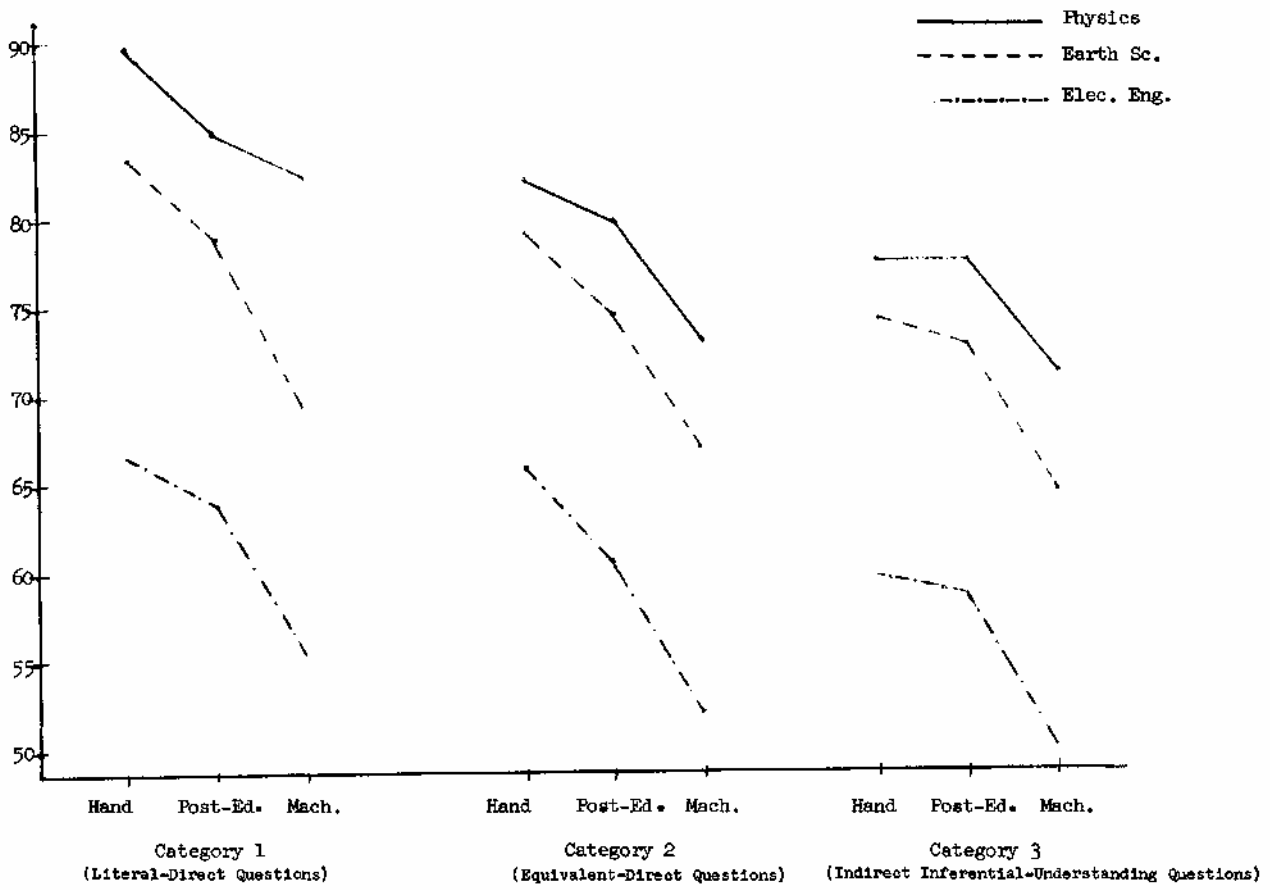

FIG. 1.—Mean percentage correct scores for three question categories for three disciplines

ADDITIONAL ANALYSES

Preliminary analyses of the linguistic characteristics of the machine translations and of the extent of input/output errors in these particular selections were carried out.

An expert translator was retained to examine the machine output in relation to the original Russian text. The analysis was designed to determine the condition leading to words completely or partially untranslated by the computer and underlined on the printouts. The conditions which may lead to an underlined word on the printout were:

1. Correct entries for which it seems reasonable that the machine should not translate them (uncommon words, proper nouns, abbreviations, etc.). There were 166 such instances in physics, 547 in earth sciences, and 224 in electrical engineering.

2. Correct entries of a common variety which should have been translated by machine, but were sometimes translated by the machine and sometimes not. There were 17 of each of such occurrences in physics and earth sciences and 103 in electrical engineering.

3. Incorrect entries in which an incorrectly spelled word or group of words were not in the computer lexicon in the incorrect form. These were printed out in full and underlined. There were 66 such errors in physics, 98 in earth sciences, and 432 in electrical engineering.

4. Incorrect entries as shown above when the word was partially translated and printed out partly in English and partly in Russian. (This also happened sometimes when there was no input error.) There were 35 such errors in physics, 57 in earth sciences, and 99 in electrical engineering.

These analyses are not reported in detail here, since it was impossible to relate them to the findings of the study in anything other than an a priori way. Suffice it to say that the considerable number of input errors found, particularly in electrical engineering, may well have reduced the comprehensibility of the machine translations to some degree.

Discussion and Conclusions

The present study has evaluated computer translations of technical Russian material from a somewhat differ-

ent point of view than that employed in the bulk of the research in this area. Comparatively little concern has been shown for traditional linguistic factors; the main emphasis has been on the communication of the technical message. Three scores were used: percentage correct answers (accuracy); total number of 10-minute time intervals to finish the test (rate); and number of items correct per 10-minute interval (accuracy rate or efficiency).

The results of the study can be summarized very briefly. With a clear and remarkable consistency from discipline to discipline and from subtest to subtest, the post-edited translation group scores were significantly lower statistically than the hand-translation group scores; and the machine-translation group scores were significantly lower than the post-edited translation group scores. The minor exceptions to the above findings that were observable on one or two subtests here and there do not impair that general conclusion. The general conclusion also holds when various types of questions are considered. If questions are categorized by type of content or questions are categorized by type of mental process involved in answering them or by directness of relationship to text or by scope of question, the same general conclusion holds.

The most important further consideration to be discussed is the extent of performance decrement. In many cases it was noted that, even though statistically significant, the difference in percentage of questions answered correctly for post-edited translations was not substantially different from that for hand translations. These simple differences were as small as 1 or 2 per cent, and, in a few instances, post-edited translations showed up as well as or better than hand translations. On the other hand, decrement for machine translation ran substantially greater. Simple differences in percentage correct ran as high as 14 per cent among the seven groups tested. Nevertheless, it should be noted that a great deal of information *was* obtainable through the machine translations. It can be hypothesized that practice in reading machine translations might improve performance on machine translations even further. There were some supporting data for this hypothesis. It is felt that in many cases machine-translation performance represented a high level of performance, even though significantly below that of the other two types of translations.

Implications for the potential improvement of the usefulness of machine translations were found in the analyses of input/output errors, linguistic analyses, and analyses of sources of inaccuracy for items with extreme differences in accuracy between hand and machine translations. These analyses indicated that in many cases the failure of the machine translation process to communicate the required information was due to input errors of one kind or another, or due to lexical errors which appeared to be correctable. If such errors were corrected, comprehension of machine-translation materials would undoubtedly rise significantly.

Although a number of interaction effects between test performance and types of material (subtests) were found, generally speaking these interaction effects were comparatively small, and it might be tentatively concluded that the findings probably apply to all types of material. It was noted, however, that there appeared to be some difference in level of performance associated with the indirectness of the content involved in the questions. In categorizing the questions into "domain" types of items, it was noted that synthesis/inference/understanding items, while producing a similar pattern of results among translation types, did so at a lower absolute level of performance than that which characterized the more direct and paraphrased items.

A further finding was the consistent suggestion that the most critical impact of using machine translations was not so much the reduction of accuracy but the increase in time (and corresponding loss in efficiency) associated with working with this type of translation. These findings were consistent with those of Pfafflin.[3] Losses on the time dimension, in terms of the percentage of decrement, were approximately double those on the accuracy dimension.

Finally, it is felt that the conclusions outlined above are quite dependable. The tests had a comparatively high degree of reliability, which was further indicated by the consistency of the observed main effects even over the comparatively short subtests. With the numbers of subjects involved, the use of the Latin Square design provided a highly powerful test for the significance of observed differences.

In closing, a word or two might be said about needed research in these areas. It will be noted that the differences between hand and post-edited translations were comparatively small. However, information external to this study suggests that the post-editing process is a very demanding and expensive process. This conclusion, in conjunction with the comparatively good overall performance of machine translations, raises the question as to whether or not training and/or practice in the use of machine-translations might be substituted for the expense involved in post-editing, with a more economical overall result. Experimentation, therefore, is needed to examine practice effects in using machine translations and to study these practice effects in conjunction with the overall cost factors associated with machine and post-editing of translations. In addition, experimentation is needed to examine the effects of varying the extensiveness of post-editing operations upon translation comprehensibility and the overall cost factors involved.

Received September 20, 1966

References

1. Edmundson, H. P. *Proceedings of the National Symposium on Machine Translation.* Englewood Cliffs, N. J.: . Prentice-Hall, Inc., 1962.
2. See, Richard. "Mechanical Translation and Related Language Research," *Science,* Vol. 144 (1964), pp. 621-26.
3. Pfafflin, Sheila M. "Evaluation of Machine Translations by Reading Comprehension Tests and Subjective Judgments," *Mechanical Translation,* Vol. 8 (1965), pp. 2-8.
4. Carroll, J. B. "Quelques Mesures Subjectives en Psycholinguistique: Fréquence des Mots, Significativité et Qualité de Traduction," *Bulletin de Psychologie,* Vol. 19 (1966), pp. 580-92.
5. Final Report on Computer Set AM/GSQ-J6(XW-2). Yorktown Heights, N. Y.: IBM, at The Thomas J. Watson Research Center, September 23, 1963. Pub. under Contract No. AF30(602)-2080; availability is limited.
6. "An Evaluation of Machine-Aided Translation Activities at F.T.D." Washington, D. C.: A. D. Little, Inc., 1965. (Available in limited quantity from A. D. Little, Inc., 1735 I St., N. W., Washington, D. C.)
7. Winer, B. J. *Statistical Principles in Experimental Design.* New York: McGraw-Hill Book Co., 1962, p. 539ff.
8. Edwards, A. L. *Experimental Design in Psychological Research.* New York: Holt, Rinehart, and Winston, Inc., 1963, p. 136ff.