TRAINING RESEARCH LABORATORY

Department of Psychology — Bureau of Educational Research
University of Illinois — 8 Lincoln Hall — Urbana, Illinois

CONTEXTUAL PREDICTABILITY AND FREQUENCY FACTORS

Domenico Parisi
Ulderico Cappelli
Lawrence M. Stolurow

Technical Report No. 41

August, 1966

Communication, Cooperation, and Negotiation
in Culturally Heterogeneous Groups

Project Supported by the

Advanced Research Projects Agency, ARPA Order No. 454
Under Office of Naval Research Contract NR 177-472, Nonr 1834(36)

Fred E. Fiedler, Lawrence M. Stolurow, and Harry C. Triandis
Principal Investigators

CONTEXTUAL PREDICTABILITY AND FREQUENCE FACTORS

Domenico Parisi
Ulderico Cappelli
Lawrence M. Stolurow

Technical Report No. 41

August, 1966

Communication, Cooperation, and Negotiation
in Culturally Heterogeneous Groups

Project Supported by the

Advanced Research Projects Agency, ARPA Order No. 454
Under Office of Naval Research Contract NR 177-472, Nonr 1834(36)

Fred E. Fiedler, Lawrence M. Stolurow, and Harry C. Triandis
Principal Investigators

Contextual Predictability and Frequency Factors

Domenico Parisi, Ulderico Cappelli,
and Lawrence M. Stolurow

Abstract

Cloze scores were obtained from 320 Ss for two written Italian passages totaling 616 words in such a way that each word was guessed by 32 Ss. Each word was classified into one of 12 grammatical classes. As has been found for English, content words are less predictable than function words if guessing the specific missing item is required. No such difference exists when only correct form class has to be predicted. Type-token ratio for each class appears to be correlated with specific item predictability, whereas proportion of occurrences of each form class in the language is correlated with form class predictability. Both correlations suggest that frequency properties may be an important factor even in complex language behavior.

Contextual Predictability and Frequency Factors

Domenico Parisi, Ulderico Cappelli
Instituto Nazionale di Psicologia, Consiglio Nazionale delle Riecerche, Rome, Italy, and
Lawrence M. Stolurow[1]

In recent years the development of psycholinguistics has fostered much interest in the long neglected grammatical aspects of language behavior. The study of the relationships among elements in linguistic sequences has been approached by a variety of techniques, mainly derived from two different and, to a large extent, opposed sources: information theory and linguists' descriptions of syntactical structure. Among the techniques inspired by information theory, statistical approximations to English have yielded a number of interesting results. However, the statistical approach is seriously limited by studying the effect of preceding context on subsequent behavior and ignoring the influence of succeeding context. Both particular studies (Goldman-Eisler, 1958; Lieberman, 1963) and general observations (Osgood and Sebeok, 1954) suggest that each linguistic segment is a function of both what precedes and follows it.

The global effect of bi-directional context can be effectively assessed by the Cloze technique developed by Taylor (1953). A number of words are canceled from a text and subjects are asked to reconstitute it by guessing the missing words. At least two dimensions of linguistic behavior can be studied by this approach: (1) predictability of a specific item and (2) predictability of the grammatical class to which the correct item belongs. Dependences among words are responsible for both lexical and grammatical predictability, but the two dimensions are partly uncorrelated and probably reflect the effects of at least two partly different determinants.

---

[1]We gratefully acknowledge the help recieved from Professor F. Agard, Department of Linguistics, Cornell University, who prepared a multi-level structural classification of the 616 words from which we selected our 12-class subdivision.

The present study has a twofold purpose. By using the Cloze procedure with two samples of Italian written language, both lexical and grammatical predictability of different form classes of Italian words will be determined and compared with data from different languages. In addition, Fillenbaum, et al., (1963) have found that in English semantic form classes (nouns, adjectives, and verbs) are more difficult to reconstitute than syntactic form classes (articles, auxiliary verbs, prepositions, and conjunctions) if scores based upon verbatim reproduction are considered, but this difference disappears when only grammatical predictability is concerned. We want to see if the same happens in Italian and, furthermore, if the relationship is influenced by varying text difficulty.

A second purpose of this study is to look for determinants of the two types of predictability. Contextual effects can be interpreted as due to long range language learning. A subject is able to predict the right word or the right form class in a particular place because of his long experience with language. Frequency has been found to be a powerful variable in rote verbal learning (Underwood and Schulz, 1960). However, the question may be asked of what effects of frequency will be when a radically different type of verbal behavior is considered. The most direct approach in assessing the relationship between frequency and contextual predictability is to use a frequency list of words such as Thorndike and Lorge have put together for English (1959). Since no such list is available for Italian, a different approach was followed which would allow the extraction of some measures of frequency of use from smaller samples of language.

Method

Materials

Two Italian prose passages (Text A and Text B) of 301 and 315 words, respectively, were used as materials for the Cloze procedure.

Text A was drawn from a daily paper and is a report of a road accident. Text B is an excerpt from a novel by V. Brancati. In order to get Cloze scores for each word in both texts, five versions of each text were prepared. Version 1 had the 1st, 6th, 11th, etc. word deleted; version 2 had the 2nd, 7th, 12th, etc. word deleted, and so on.

Subjects

320 students of 17 to 22 years of age were used as subjects. About one-half were male and one-half female. One-third of the sample were students in the last year of high school, and the remaining two-thirds were college students.

Procedure

Ss were randomly given one of the two mutilated texts with instructions to fill in all the blanks with the words they thought most likely to appear in the intact text. Each S had one of the five versions of either Text A or Text B. Therefore, each of the 616 words was guessed by 32 Ss. Time for completing the work was unlimited, but Ss were told in the instructions that they should finish in about 10 or 15 minutes.

Results

For each word in the two passages a verbatim (V) score and a form class (FC) score were computed. V score was percentage of Ss filling in the blank with a word either identical to the missing word or just clearly misspelled. A FC score was the percentage of Ss giving a word which was in the same grammatical class as the correct word. Mean V score was 67 per cent for Text A and 54 per cent for Text B. This difference was taken as a difference in text difficulty (Taylor, 1953).

Each of the 616 words of the two texts was classified into one of 12 grammatical classes: nouns (N), qualifying adjectives (ADJ), verbs (V), adverbs (ADV), quantitative adjectives (Q), articles (AR), prepositions (PRE), conjunctions (C), auxiliary verbs (AV), pronouns (P), other adjectives (OA), and non-classified (NC). Table 1 shows number of items and V and FC scores for each

grammatical class, both for each text separately and for both texts together. Also shown are V and FC scores for content words (nouns, qualifying adjectives, verbs, adverbs, quantitative adjectives) and for function words (the remaining ones). If guessing of specific items is required, content words are more difficult to reconstitute than function ones. If only form class is considered, the difference disappears. Both results are in agreement with findings reported for English by Fillenbaum et al. (1963). Furthermore, the difference in specific item difficulty between content and function words appears to increase with text difficulty, as it obviously should, since text difficulty depends much more on content word difficulty than on function word difficulty. On the other hand, if FC scores are considered, the two texts do not differ very much in either overall difficulty or differential difficulty of content and functional items.

Fillenbaum et al. (1963; see also Ervin-Tripp, and Slobin, 1966) attributed the differential predictability of various grammatical classes to class size, that is, to the number of items included in each class. To verify this hypothesis, the rank order correlation coefficient between verbatim predictability and number of different items in each class in our 616 word sample was calculated. This coefficient is -.21, which is well-below significance. The conclusion that predictability of specific words in a context is determined not by the size of the class they belong to, but by the type-token ratio of that class, which is an index of frequency of use, seems, therefore, to be warranted.

The type-token ratio (TTR) was calculated for each grammatical class on the basis of both Texts A and B as language sample. That is, for each of the 12 classes the number of different items occurring in both texts was divided by the total number of items in the class. The rank-order correlation between TTR and mean V score for each class was -.80, which is significant at $p < .01$ (Figure 1). Furthermore, the number of occurrences of each class

Table 1

Number of Items, V and FC Scores for Each Grammatical Class

Text	Classes	N	ADJ	V	ADV	C	AR	PRE	C	AV	P	OA	NC	Words Content	Function	Total
	Number of items	80	17	33	12	10	24	58	14	31	10	5	7	152	149	301
A	V	32	56	55	38	57	86	83	82	64	80	44	33	58	76	67
	FC	97	83	92	83	60	92	96	85	95	82	64	55	91	90	90
	Number of items	75	44	22	10	2	33	61	19	13	20	12	4	153	162	315
B	V	39	29	37	40	47	75	75	53	74	76	67	65	36	71	54
	FC	93	83	83	82	47	83	92	72	89	86	71	77	87	85	86
	Number of items	155	61	55	22	12	57	119	33	44	30	17	11	305	311	616
Both	V	51	36	48	39	55	80	79	65	67	77	60	44	47	74	60
	FC	95	83	88	83	65	87	94	77	93	85	69	63	89	87	88

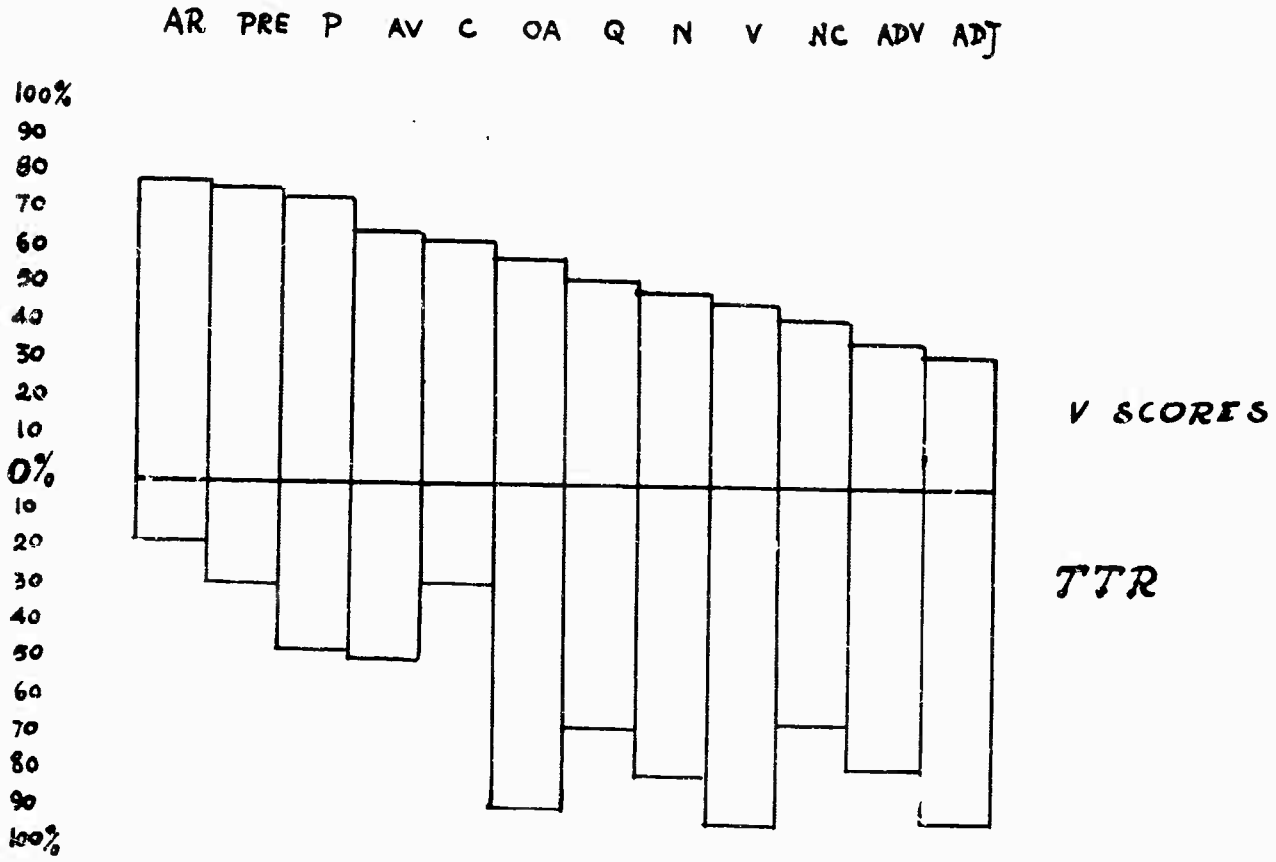

FIGURE 1. RANK-ORDER CORRELATION BETWEEN V SCORES AN TTR FOR EACH GRAMMATICAL CLASS ($r = -.80$)

was correlated with the mean FC score of that class, and the rank-order correlation coefficient was +.86 which is significant at $p < .001$ (Figure 2).

Discussion

The results of the present study show that predictability properties of written passages are remarkably homogeneous across languages. When conditions are similar, as in this study and in deletion rate five of Fillenbaum, et al. (1963), the order of verbatim difficulty of content classes appears to be the same for Italian and English: qualifying adjectives, adverbs, verbs, nouns, and quantitative adjectives. Differences in functional classes may be due to discrepancies in grammatical classification. More generally, in both Italian and English (Fillenbaum, et al., 1963; Aborn et al., 1959; Coleman and Blumenfeld, 1963) predicting that a word is a content or a functional item is about as difficult, but differential difficulty shows up when one is asked to predict the specific content or functional item. Both in Italian and English content words are twice as difficult as function words.

Classification in content or functional classes is very broad. More specific determinants of predictability can be found by searching through the frequency properties of language. Type-token ratio of a particular grammatical class can be used as an index of the mean frequency of use of a type in that class. Out of 100 nouns actually used, 92 are different nouns. Mean frequency for a noun type is 1.09. Out of 100 articles actually used, only 16 are different words. Mean frequency for articles is 6.25. These frequency properties of grammatical classes appear to determine to a remarkable degree the mean predictability of words in each class. The predictability of a particular word in a text is a function of type frequency of the grammatical class to which it belongs.

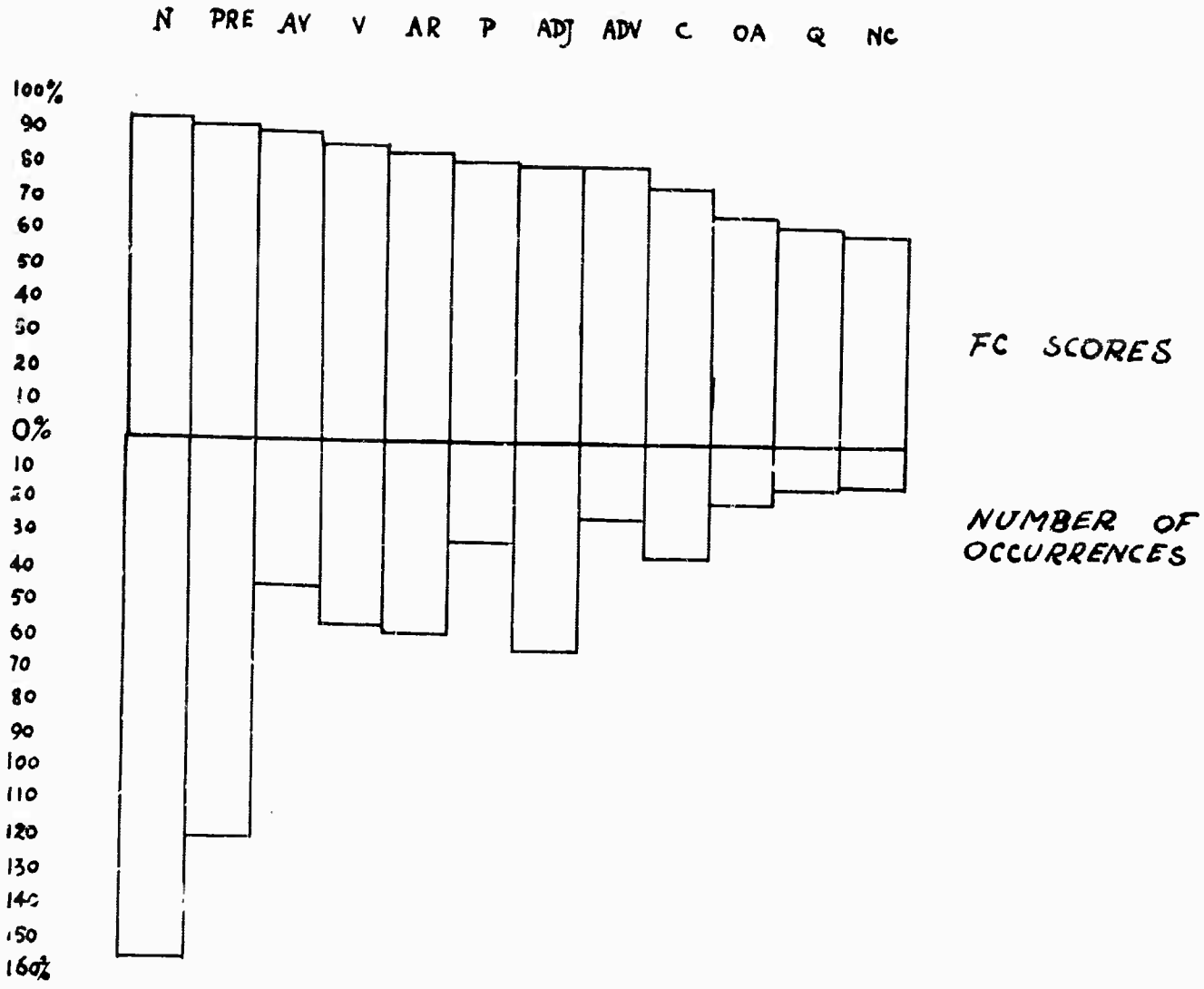

FIGURE 2. RANK-ORDER CORRELATION BETWEEN FC SCORES AND NUMBER OF OCCURRENCES FOR EACH GRAMMATICAL CLASS ($\tau = +.86$)

A more straightforward relationship between frequency and predictability may be seen in the form class data. Here, guessing the right grammatical class in a particular context appears to be a function of frequency of use of that grammatical class in the language.

Apart from the obvious limitations of the present study, a general conclusion can be drawn regarding the determinants of complex linguistic behavior. Guessing the right word or the right grammatical class in a context seems to be a complex task in which all sorts of sequential, both syntactic and semantic, cues should influence performance. It is because these more complex determinants of behavior are absent in most experiments involving verbal material that frequency may emerge as an important factor in simpler tasks such as rote verbal learning tasks, word recognition tasks, and so on. The present data, however, seem to show that frequency may be an important factor even in a linguistically very sophisticated task such as predicting words in a context. More specifically, it could require an extension of the "spew" hypothesis put forward by Underwood and Schulz (1960), in which frequency of experience with a particular verbal unit determines its availability, to complex verbal behavior.

References

Aborn, M., Rubenstein, H., and Sterling, T. D. Sources of contextual constraint upon words in sentences. Journal of Experimental Psychology, 1959, 57, 171-180.

Coleman, E. B. and Blumenfeld, J. P., Cloze scores for nominalizations and their grammatical transformations using active verbs. Psychological Reports, 1963, 13, 651-654.

Ervin-Tripp, S. M. and Slobin, D. I. Psycholinguistics. In Annual Review of Psychology, 1966, Palo Alto, California.

Fillenbaum, S., Jones, L. V., Rapoport, A., The predictability of words and their grammatical classes as a function of rate of deletion from a speech transcript, Journal of Verbal Learning and Verbal Behavior, 1963, 2, 186-194.

Goldman-Eisler, Frieda, Speech production and predictability of words in context. Quarterly Journal of Experimental Psychology, 1958, 10, 96-106.

Lieberman, P., Some effects of semantic and grammatical context on the production and perception of speech. Language and Speech, 1963, 6, 172-187.

Osgood, C. E. and Sebeok, T. A. (Eds.). Psycholinguistics: A Survey of Theory and Research Problems. Baltimore: Waverly Press, 1954.

Taylor, W. L., Cloze Procedure: A new tool for measuring readability. Journalism Quarterly, 1953, 30, 415-433.

Thorndike, E. L. and Lorge, I. The Teacher's Word Book of 30,000 Words. New York: Bureau of Publication, Teachers College, Columbia University, 1959.

Underwood, B. T. and Schulz, R. W. Meaningfulness and Verbal Learning. Chicago; Lippincott, 1960.

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)
University of Illinois

2a REPORT SECURITY CLASSIFICATION

2b GROUP

3 REPORT TITLE
Contextual Predictability and Frequency Factors

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report

5 AUTHOR(S) (Last name, first name, initial)
Parisi, D., Cappelli, U., and Stolurow, L. M.

6 REPORT DATE	7a TOTAL NO. OF PAGES	7b NO OF REFS
August, 1966	10	10

8a CONTRACT OR GRANT NO	9a ORIGINATOR'S REPORT NUMBER(S)
NR 177-472 Nonr 1834-36	41
b PROJECT NO	9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10 AVAILABILITY/LIMITATION NOTICES

11 SUPPLEMENTARY NOTES

12 SPONSORING MILITARY ACTIVITY
Office of Naval Research

13 ABSTRACT

Cloze scores were obtained from 320 Ss for two written Italian passages totaling 616 words in such a way that each word was guessed by 32 Ss. Each word was classified into one of 12 grammatical classes. As has been found for English, content words are less predictable than function words if guessing the specific missing item is required. No such difference exists when only correct form class has to be predicted. Type token ratio for each class appears to be correlated with specific item pr dictability, whereas proportion of occurrences of each form class in the language is correlated with form class predictability. Both correlations suggest that frequency properties may be an important factor evn in complex language behavior.

DD FORM 1 JAN 64 1473 0101-807-6800

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Cloze scores function words content words specific item predictability form class predictability frequency properties complex language behavior						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.

DD FORM 1 JAN 64 1473 (BACK)