# English Article Insertion*

by Jocelyn Brewer, Colorado State University, Fort Collins

*For an 8,300-word sample of English text we have found that it is possible to provide at least an acceptable article for more than 90 per cent of the noun occurrences at a "cost" of providing a dual article for half of the occurrences. This can be achieved by making use of the following relatively simple criteria for article selection: (1) prior classification of nouns according to the articles they are expected to take in natural-language text, (2) grammatical number of the noun, (3) presence or absence of a following "of" phrase, and (4) presence or absence of certain specified modifiers. A study of noun classification indicates that it can be done with acceptable consistency and reliability. The recommended pattern of article insertion was implemented as part of the Bunker-Ramo machine-translation program and tested on a brief sample text. This work has indicated that a certain amount of further improvement in article insertion can be achieved by extension of the above criteria but that further progress will require dealing with articles on the semantic level—in terms of semantic attributes and semantic relations.*

## Introduction

Although to a very considerable extent English articles are determined by context, both within and beyond the boundaries of the sentence in which they occur, and hence may be considered semantically redundant, they are so basic a part of idiomatic English that their absence from a machine-translation output results in a product that is linguistically extremely unpalatable. When translating from a language without articles, such as Russian, there is in some cases no indication as to which article would have been appropriate to the intent of the author. However, we should like to be able to exploit all the contextual clues that do exist. These are found generally to be of a semantic rather than syntactic nature. Since the present machine-translation program relies primarily on syntactic analysis and is not yet prepared to deal with all the semantic complexities of natural language, we should like at this time to isolate and identify in its simplest form that kind of semantic information which specifically bears on the problem of article usage and which represents the minimum that must be supplied to allow for acceptable article insertion.

This is a somewhat different problem from a general analysis of article function, such as that undertaken from a transformationalist point of view by Beverly Robbins and others at the University of Pennsylvania, although the partial analysis required for machine translation must be reconcilable with a more general

theory. The general analysis of article function can take as data such linguistic elements as intonation and punctuation, and indeed must analyze the nuances of meaning that articles are used to express. But in machine translation the problem is to generate these, given only the source-language text, as rendered into machine-readable form, and such syntactic and semantic tags as may be attached to the forms that occur. The problem is then to manipulate these elements in such a way as to reflect the meaning equivalences between source and target languages and to comply with the requirements of natural-language usage. It is neither necessary nor at this time possible to exploit all the English patterns that are available to the native speaker of English.

This study represents an attempt to discriminate between elements of the article-insertion problem that are amenable in a practical way to semantic resolution and those that should better be dealt with on a statistical basis related to observed frequency of occurrence in text. In an earlier study by Martins [1] a method of article insertion was proposed which was intended to produce an acceptable machine-translation output, without necessarily duplicating the articles used in any given text. In brief, it was proposed: (1) to recognize three articles: "the," "a/an," and "0" (no explicit article); (2) to classify nouns in the machine-translation dictionary into six classes for purposes of article insertion; (3) to apply the dual syntactic criteria of *(a)* whether singular or plural and *(b)* whether followed by a linked genitive block or not in order to further limit the articles to be supplied to one or, at most, two; (4) to print both article choices when there are two, omitting the "0" article designation only when it is the only choice; and (5) to omit any article when

a noun is preceded by any of a specified list of modifiers.

In Section I we report on a study of noun classification. In Section II we present the results of a detailed analysis of the distribution of articles and their inter-substitutability in the sample text, recommend a somewhat modified article-insertion pattern on the basis of this study, and discuss some of the mechanisms that appear to account for the observed pattern of article use. In Section III we evaluate the article insertion in a machine-translation output that resulted from incorporating the basic recommendations into the Bunker-Ramo machine-translation program.

The sample text selected for analysis comprised three English articles totaling approximately 8,300 words, all dealing with some aspect of language translation in order to insure some overlap in vocabulary: (1) H. Wallace Sinaiko, "Experiment in International Teleconferencing," 1,600 words; (2) Edgar Hammond, "Traduttore, Traditore," *International Science and Technology* (October, 1962), 3,100 words; (3) Gilbert W. King and Hsien-Wu Chang, "Machine Translation of Chinese," *Scientific American* (June, 1962), 3,500 words. For evaluation of the article-insertion scheme in our machine-translation program we used a machine translation into English from a Russian version of the same article by Sinaiko, which had originally been prepared for the purpose of obtaining comparable translations from various machine-translation groups.

## I.   Study of Noun Classification

The article-insertion scheme of Reference 1 had established six noun classes (five, plus the category of nouns that never take an article) for purposes of article insertion, and we wished to verify their validity as discrete and stable categories. Further, the scheme provided for assigning both the singular and the plural forms of a noun to a single class, depending upon criteria applied to the singular form alone. We wished to determine whether a single article prescription was consistently appropriate to all plural forms of the nouns that had been placed in the same class on the basis of tests applied to the singular forms only. A further problem was that no procedure had been provided for classifying those nouns for which there is no singular form. And finally we wished to test the operational feasibility of the proposed classification procedure.

### A.  CODING OF NOUNS OUT OF CONTEXT

This phase of the study was conducted without reference to the articles actually occurring with these nouns in the text. A total of 710 nouns, including certain pronouns that may on occasion take articles, were recorded from the three articles of the sample text. The entire group of nouns was coded twice and the results compared for consistency.    The first classification was

carried out by simply testing the intuitive acceptability of "the," "a/an," and "0" in turn with each noun. Singular and plural forms were classified independently and coded according to the following:

| Acceptable Articles | Letter Code |
|---|---|
| the, a, 0 | A |
| the, a | B |
| a, 0 | C |
| the, 0 | D |
| the | E |
| a | F |
| 0 | G |

For example, the word "table" was assigned to class B on the basis of finding it acceptable to talk about "a table2 or "the table," but rejecting "(0) table" without an explicit article. The word "supervision" was assigned to class D on the basis of accepting the combinations "the supervision" and "(0) supervision" and rejecting as unlikely "a supervision." Classes C and F were found to be empty.

Then the entire group of nouns was reclassified in accord with the coding procedure proposed in Reference 1 (the classes being here renumbered from 1 to 6 for ease of reference):

0. Is the noun always used without an article?
    Yes:  Class 6
    No:   See rule 1

1. Can the noun, in the singular, begin a sentence of the type: "——— is necessary," etc.?
    Yes:  Class 3
    No:   Class 5

3. Does this noun, in the singular, *always* require "the"?
    Yes:  Class 4
    No:   See rule 4

4. Is the meaning of this noun intuitively more abstract than concrete, or is its meaning vague?
    Yes:  Class 2, tentatively
    No:   Class 1

The essential equivalence between the two sets of classes is shown in Table 1.

TABLE 1

| Criterion | Numerical Code | Possible Articles | Equivalent Letter Code |
|---|---|---|---|
| Never an article....... | 6 | 0 | G |
| Sometimes "0" article: | | | |
|   Never "a"............ | 5 | The, 0 | D |
|   Any ................... | 3 | The, a, 0 | A |
| Always an article: | | | |
|   Always "the" ...... | 4 | The | E |
| Noun is abstract or | | | |
|   vague ........... | 2 | The, a | B |
| Noun is not abstract | | | |
|   or vague .......... | 1 | The, a | B |

Comparison of the results of the two classification procedures showed a high degree of consistency between the class assignments and appeared to confirm the stability of the categories. The discrepancies with respect to classification of singular nouns all involved classes 1 and 2, where, of the 352 nouns assigned to these classes by the numerical coding procedure, 38 had been given the less restrictive letter code A, which allows for all three possible articles. This reflects the fact that for some nouns for which it is not acceptable to say "——— is necessary" other contexts were created in which the noun was expected to be used without an explicit (with the "0") article. The numbers of nouns assigned to the various numerical classes are shown in Table 2.

TABLE 2

| Class | Number |
| --- | --- |
| 1 ............................................................... | 314 |
| 2 ............................................................... | 38 |
| 3 ............................................................... | 250 |
| 4 ............................................................... | 26 |
| 5 ............................................................... | 52 |
| 6 ............................................................... | 23 |
| Uncoded (no singular form) ........................ | 7 |
| Total   .................................................... | 710 |

It was found that for nearly all nouns for which a plural form exists, either "the" or "02 was considered possible, regardless of the classification of the singular form. For the 116 of the 710 nouns for which a plural form was not believed likely, any article prescription for plural forms would simply not be applied. It was found that plural forms usually exist for nouns of classes 1, 2, and 3 but are rare for nouns of classes 4, 5, and 6. Hence a single class, "plural" is proposed for most plural nouns, regardless of the classification of the singular form.

There were, however, seven plural nouns for which only the article "the" was expected: "Japanese," "Chinese," "English," "Spanish," "French," "hallmarks," and "contents." Five of these are names of nationalities which are, in fact, not plurals of the singular form; these refer to the language when used in the singular without an article but refer to people when used in the plural. It would be desirable to establish a class for such plurals for use with "the" only. Only a single plural form was encountered that can occur with "the," "a," and "0"—the anomalous pronoun "few," which may be used with all three, with marked differences in meaning. (Other collective nouns, such as "group," can be classified regularly as singular forms.)

B.   CODING PROCEDURE

The greatest difficulties in coding arose in *(a)* applying the criterion of "vagueness" or "ambiguity" to separate class 2 from class 1 nouns and *(b)* applying a single code to nouns with multiple meanings. Since the ratios between the uses of "the" and "a" for singular and "the" and "0" for plural occurrences of the nouns of the two classes were approximately the same, and since the separating criterion does not seem sufficiently clear to be operationally effective, class 2 was assimilated into class 1, thereby reducing the number of classes for singular nouns to the five that represent the actual article combinations found to occur. They will be identified hereafter as follows: class 1: "the," "a"; class 3: "the," "a," "0"; class 4: "the"; class 5: "the," "0"; class 6; "0"

Nouns with multiple meanings were dealt with summarily by assigning a code sufficiently broad to include the appropriate articles for all anticipated meanings of each noun. This resulted in assigning many words to class 3 when the separate meanings could have been assigned to classes 1, 5, or 6.

A rather sensitive method for revealing the existence of multiple meanings represented by a single noun form, each alone taking a more narrow article code, involves testing each noun with the modifier "such." The following combinations are found to occur:

| | | |
| --- | --- | --- |
| Class 1 | Only "such a——" : | "Such a chairman," "such a group" |
| Class 3 | Both, *if* the noun's meaning changes when "such" is replaced by "such a": | |
| | "Such a——" | Class 1-type meaning: "Such a language," "such a communication," "such a German" |
| | "Such——" | Class 5- or 6-type meaning: "Such language," "such communication," "such German" |
| Class 4 | Neither: | Class 4 nouns would not normally be used with "such": "Upshot," "worst," "Andes," "beautiful" |
| Class 5 | Only "such——": | "Such clothing," "such information," "such transportation" |
| | Or both, *if* the noun's meaning does *not* change when "such" is replaced by "such a": | "Such oil" "such an oil," "such appreciation ≈ such an appreciation," "such sympathy ≈ such a sympathy" |

Class 6  Rarely either:

Class 6 nouns would rarely be used with any article and are very rarely used with "such":
"Such a Europe," "such a mankind," "such plenty"

The following classification routine is based on these findings (an appropriate modifier may be placed before the noun):

1. Would you expect the noun to be used with "the" or "a/an"?

     No:   Class 6
     Yes:  Go to 2

2. Can one say "such a——"?

     Yes:  Go to 3
     No:   Go to 5

3. Can one also say "such——"?

     Yes:  Go to 6
     No:   Go to 4

4. Would you expect the noun to be used without (with the "0") an article?

     No:   Class 1
     Yes:  Class 3. Go to 8

5. Can one say "such——"?

     Yes:  Class 5
     No:   Class 4

6. Are the meanings with "such" and "such a" the same?

     Yes:  Class 5
     No:   Class 3. Go to 7

7. The meaning with "such a" is a class 1-type meaning. Using the meaning of the noun with "such," would you expect to say "the——"?

     Yes:  Class 5-type meaning
     No:   Class 6-type meaning

8. The meaning with "such a" is a class 1-type meaning. The meaning when the noun is used without an article is a class 6-type meaning.

Unfortunately, though semantic criteria are at hand to classify the various meanings of the class 3 nouns, machine-recognizable criteria are difficult to define. Hence class 3 is being retained at present for machine-translation purposes.

It is found that the coding of nouns out of context proceeds rather rapidly by whatever procedure. When coding, it soon becomes clear that for most nouns one can create contexts using any of the three articles and that the classification actually represents, in many if not all cases, a statement of expectation rather than a description of the only possibilities. Nonetheless, judgments as to the likely articles seem sufficiently consistent to serve the present purpose.

C.  NOUN CHARACTERISTICS BY CLASS

In order to interpret the significance of this kind of classification, let us consider the common characteristics of the nouns assigned to each of the article classes. In brief:

*Class 1.*—The noun referents are found to be enumerable or to occur as discrete entities: "the/a table," "the/a problem," "the/a group."

*Class 3.*—These nouns may be used either with a class 1-type meaning (i.e., referring to discrete or enumerable entities) or with a class 5- or class 6-type meaning. The meanings may or may not be similar, although often the class 5- or class 6-type meaning is an abstraction or a generic term and the class 1-type meaning a discrete embodiment of it. Compare "the/a necessity" with "the/0 necessity," "the/a translation" with "the/0 translation," "the/a case" with "the/0 case," "the/a Italian" with "(0) Italian," "the/a duty" with "(0) duty," "the/a man" with "(0) man."

*Class 4.*—This class appears to include at least three subgroups: (1) superlatives and nouns and pronouns whose referent is completely determined in a given context, as "the best," "the like," "the outset," "the upshot"; (2) adjectives used as generic nouns, as "the beautiful," "the disenchanted"; and (3) those proper nouns which require "the": "the Andes," "the *Herald Tribune,*" "the United Nations," "the Tigris."

*Class 5.*—The referents are abstract or generic. They include abstract entities, qualities, processes, attributes, and generic names for matter, as "praise," "information," "guesswork," "transportation," "sand," "oil," and most gerunds: "thinking," "decoding."

*Class 6.*—This class again appears to include two subgroups: (1) The first includes rarely modified nouns such as "mankind" and "womanhood," which can be forced to take an article only with difficulty. (2) The second includes most proper names, as "Europe," "IBM," "Y. R. Chao."

Let us now consider these groups in more detail. With the singular class 1 nouns, the required article, whether it be "the" or "a," appears to carry a double burden. The feeling that some explicit article is needed reflects an awareness that the referent of the noun is discrete and enumerable. That is, the article, *qua* article, corroborates the class 1 characteristics of the noun referent. Further, the article may denote particularity or non-particularity according to the context (including punctuation in written and intonation in spoken language). In those cases where either article is appropriate, either where a generic meaning of "the" coincides with the "representative sample" meaning of "a" or where the noun referent is sufficiently narrowly identified by modifiers in context as to narrow the possibility of interpretation to one, *some* explicit article is still required to serve the first purpose, even though the articles may be substitutable.

Class 3 nouns are identified by the coding procedure as those that may take any of the three articles. The coding procedure based on a test frame of "such" will usually serve to identify the appropriate article classes of the different meanings represented by a noun. Although it was sometimes easier to assign more restrictive article codes when a noun was considered in isolation than when embedded in "live" text, thereby revealing the somewhat artificial and procrustean nature of the present five classes, for the greater number of occurrences of class 3 nouns the distinction *is* clear. In general the referents of the class 1-type meanings are, as for class 1 nouns, discrete and enumerable and often concrete. The referents of the class 5-type meanings, like those of the class 5 nouns, are generic, non-enumerable, and often abstract. In general the referents of the class 6-type meanings are highly abstract, and "the" cannot even be used generically with them without changing their sense, as with "duty" and "man."

The referents of class 4 nouns, which are expected always to occur with "the," appear to be semantically restricted either to particularity (the superlatives, proper nouns, and those nouns that are restricted to a single referent in any given context) or to generality (adjectives used as nouns). For the proper nouns in this class that require the double indication of particularity, capitalization *and* the definite article, this redundancy may be regarded as an idiomatic requirement. Perhaps, however, it is no accident that this pattern is generally required for rivers, oceans, and mountain ranges, which are certainly less bounded, metaphorically speaking, than lakes, mountain peaks, and cities.

*Class 5 nouns.*—The very nature of their referents is non-discrete. One may say in general that they can be particularized in meaning but not enumerated. For example, one may speak of "information" in general, or of "the information," but it cannot be counted. Except with the mass nouns ("the wind," "the water," "the snow"), "the" is seldom used generically. When "the" is used with class 5 nouns it usually means "some particular." The only open issue relevant to article use is particularity versus generality. We find that "the" is usually *required* only when it is necessary to denote particularity explicitly; "0" is *required* only when it is necessary to denote non-particularity or generality. As with plural nouns, we find that, when particularity is clearly implied by the context, "the" may be used but is often not required, and economy of wording appears often to result in a preference for "0."

It is true that class 5 nouns may be used with "a," as in the phrases "arose from *an* early *recognition*," "need for *a* stringent *formalization*," "acceptance that *a* real *translation* is impossible," "he felt *a* deep *anxiety*," "*a* very fine *sand*," but we propose to omit this alternative for machine translation. These may be con-

sidered as elliptical constructions in which "a" introduces the idea "kind of" explicitly or implicitly; its use is usually optional, the more prosaic "0" being substitutable for it with little change in meaning. Class 3 nouns may be distinguished from those of class 5 by the fact that the meaning of the word when used with "a" (the class 1-type meaning) is clearly different from its meaning when used with the "0" article, as with "a communication" versus "communication." For class 5 nouns no change in meaning results from changing the article, as with "a sympathy" versus "sympathy," or "an intensity" versus "intensity."

The two subgroups of class 6 nouns appear to require the "0" article for different reasons. The referents of the abstract nouns are generally understood to be neither discrete nor enumerable; hence, no article is required to establish the presence or absence of these attributes. The proper names of class 6 are semantically akin to class 1 nouns in that their referents are discrete and enumerable. When the device of capitalization is sufficient to indicate particularity, no article is required. Conversely, when no article is used, the particularity of a proper noun is understood if the noun can be so construed. Consider the differences between (1) a fully specified name, such as "Gilbert W. King," which requires no article; (2) a proper noun which is nonetheless used in a non-restricted sense, as in "There is a red-headed Gilbert in the class"; and (3) "King taught the class," where absence of article *denotes* the particularity of a proper noun.

With plural nouns, their very plurality generally indicates that the referents are discrete and, ipso facto, enumerable. This is why plurals of class 3 nouns are plural forms of their class 1-type meanings. The plurals of the names of nationalities are semantically no different from other plurals, but, when there is no orthographic change from the singular form to the plural, it appears that a different noun form is required with the indefinite article to avoid ambiguity. Hence, we have "French," singular, a class 6-type meaning, and "the French" or "(0) Frenchmen," plurals of the class 1-type meaning.

In contrast to the situation with class 1 nouns, for plural nouns the article only serves the second article function. Often "the" is only *required* if it is necessary to establish particularity, and "0" is only *required* if it is necessary to establish non-particularity. As with class 5 nouns, when the issue is not important, usually because the meaning is implicit in the context, use of "the" may be optional and no explicit article required.


## II.   Article Use in the Sample Text

In a second phase of this study we turned to the actual article distribution in the three articles of the sample text in order to evaluate the noun-coding and proposed article-insertion scheme and to derive further rules for

more precise article insertion. We wished in particular to investigate: (1) the number and nature of exceptions in the English text to the articles designated by our coding of the nouns out of context, (2) the extent to which the articles used in the sample text were supplied by the proposed article-insertion scheme, (3) in how many of the cases in which the proposed article-insertion scheme failed to supply the article used in the sample text the article that was supplied was still acceptable, and (4) the relation between the number of articles allowed by noun-coding, the number supplied by the article-insertion scheme, and the number of acceptable insertions. An extremely careful study was done of the intersubstitutability of the articles in the sample text in order to estimate the tradeoff between omitting certain of the articles anticipated on the basis of the noun-coding and the errors that would result. Finally we attempted to extend the number of instances in which we could specify articles in terms of context more precisely than by coding alone.

## A.   ANALYSIS OF ARTICLE DISTRIBUTION

First we wished to obtain a count of the article occurrences in the sample text, grouped by article class of the noun, by number, and by presence or absence of a following genitive phrase. However, for a number of noun occurrences, the article (or its absence) is dictated by elements of context that override the normal article usage. For example, certain preceding modifiers, such as "some," "any," "no," etc., suppress, or replace, any article. In such cases, the article was considered non-existent and not counted as a "0" article. Nouns are commonly used without articles in short titles and headings; these, too, were excluded from our count. Also, occurrence in an idiom frequently dictates an article usage not otherwise typical of a noun, and so obvious English idioms were excluded from the count. With these exceptions, the nouns of the three articles of the sample text were listed with the accompanying article, "the," "a/an," or "0," and sorted according to article class, whether singular or plural and whether or not followed by a modifying "of" phrase (the English equivalent of the "syntactically linked genitive block" of the machine-translation syntactic-analysis program). Since the modifier "one," when used without "the," substitutes for "a/an," all such occurrences were included in the count for "a/an."

Of the 1,027 occurrences of singular nouns that were considered, there were 29 instances of articles occurring (in each case, the "0" article) that were not compatible with the classes to which the nouns had been assigned. Of these 29, 20 occurred in idioms that had been overlooked in error, 2 instances were deemed to represent exceptional usage, and 7 appeared to be candidates for transfer from class 1, which excludes the "0" article, to class 3, which allows for it. This is indeed a small number of exceptions to noun-coding done

without reference to the context from which the nouns were taken, and definitely confirms the feasibility of at least restricting the articles to be inserted to those that are compatible with the article coding of the nouns.

On the basis of *classification alone,* multiple article possibilities were recognized for most of these noun occurrences of the sample text (Table 3). The article-

TABLE 3

| No. of Articles | No. of Noun Occurrences | Percentage |
|---|---|---|
| 0 ("0")   ............................. | 72 | 5 |
| 1 ("the")   ......................... | 20 | 1 |
| 2 ("the/a" or "the/0")   ........ | 1,063 | 69 |
| 3 ("the/a/0")   ..................... | 378 | 25 |
| Total   ............................... | 1,533 | 100 |

insertion scheme proposed in Reference 1 would omit certain articles allowed by the noun-coding in the interest of reducing the number of multiple articles to be supplied. The articles prescribed by this scheme were compared with those occurring in the sample text. In each class where it was attempted to eliminate one of the articles allowed by the noun-coding there were exceptions. Since, however, it was the intent to provide an acceptable English reading rather than to duplicate the articles actually used, the exceptions were listed in context and scored according to whether or not the proposed article or at least one of the alternatives provided would have allowed for an acceptable reading. Any resultant change in meaning was not taken into account, except insofar as the wider context dictated a specific meaning which the article would have to express.

For the occurrences of the 483 nouns in those classes where an article allowed by the coding had been excluded, 126, or approximately one-fourth, were not provided with the same article used in the text. Of this fourth, approximately 55 per cent of the insertions were nonetheless acceptable and 45 per cent were not. In terms of text as it would have appeared to the reader, with articles supplied in accordance with this scheme, the results were as shown in Table 4. In

TABLE 4

| No. of Articles Supplied | No. of Noun Occurrences | No. of Unacceptable Insertions | Percentage of Occurrences Unacceptable |
|---|---|---|---|
| 0 ("0")   ........... | 122 | 0 | 0 |
| L ("the")   ...... | 77 | 15 | 1 |
| 2 ("the/a" or "the/0")   . . . | 1,334 | 42 | 3 |
| Total   ....... | 1,533 | 57 | 4 |

summary, providing dual articles to seven-eights of the nouns resulted in 4 per cent unacceptable insertions.

It is seen that, in comparison to the articles provided on the basis of noun-coding alone, the number of noun occurrences with a single article is about double; the occurrences coded for three possible articles have been restricted to two of the alternatives. These figures are more revealing when expressed in terms of articles omitted (Table 5). In other words, of these

TABLE 5

| No. of Possible Articles Omitted | No. of Occurrences | No. Unacceptable |
|---|---|---|
| 0 .................. | 1,050 | 0 |
| 1 .................. | 483 | 57 |
| Total ......... | 1,533 | 57 |

noun occurrences (excluding idioms and those situations in which the article use was clearly determined) less than 4 per cent of the total insertions (57 out of 1,533) failed to include an acceptable article; But, when only that group of occurrences is considered where a possible article was omitted, approximately one out of eight (57 out of 483) was not provided with an acceptable article. It became apparent that to determine the optimum limit of multiple-article reduction it would be necessary to know the tradeoff between reducing the number of multiple articles inserted and failing to provide an acceptable article.

### B. ANALYSIS OF INTERSUBSTITUTABILITY OF ARTICLES IN THE SAMPLE TEXT

To this end a careful and exhaustive study was undertaken to determine the extent to which articles are substitutable, one for another, with respect to nouns of each class. It was attempted to account for every noun of the sample text, excluding only passages in quotation marks that were not intended to represent natural English usage. Nouns in idiomatic occurrences, proper names, and titles were included. 1,710 noun occurrences were examined; the 255 additional occurrences where the article was suppressed by a preceding modifier were noted but did not enter further into the analysis.

For every noun occurrence, each article ("the," "a," and "0") was tested for acceptability in that particular context. Numbers written out in words were included. A record was made of the article actually used and any acceptable substitute(s). After these data had been recorded for each noun, its article class was looked up in the coding file and added to the record. The class distribution is shown in Table 6.

Analysis of the results showed that for class 1 singu-

TABLE 6

| CLASS | NUMBER | |
|---|---|---|
| | Singular | Plural |
| 1 ........................................... | 537 | 345 |
| 3 ........................................... | 426 | 242 |
| 4 ........................................... | 22 | 0 |
| 5 ........................................... | 47 | 1* |
| 6 ........................................... | 79 | 2† |
| Plural form only ........................................... | | 9‡ |
| Total ................................... | 1,111 | 599 |

Total coded ........................................... 1,710
Occurrences with article suppressed........................ 255

Total noun occurrences ....................................... 1,965

* "Negotiations."
† "The French," "(0) plenty of . . ."
‡ "(0) people"—four occurrences; "the people"—two occurrences; "(0) seven-eighths of . . ."; "(0) two-thirds of . . ."; "(0) auspices."

lar nouns the presence of a following "of" phrase did not appear to affect article selection. The article "the" was used for 53 per cent of the occurrences and would have served for another 7 per cent. The article "a" was used for 40 per cent of the occurrences and would have served for another 17 per cent. The "0" article was used for 7 per cent of the occurrences, all of which were considered to be idiomatic or to represent exceptional usage. Supplying the best single article, "the," would have resulted in 40 per cent unacceptable insertions for this group.

The figures for the occurrences of class 3 singular nouns substantiate the premise that this group is comprised of nouns with multiple meanings. For only 9 out of the 426 occurrences did all three articles appear to be acceptable. In each of these cases there was only a trivial difference in meaning among the three article possibilities, and the noun could have been assigned to class 5. For an additional 20 out of the 426 occurrences, "a" and "0" were recorded as alternately acceptable. In some of these occurrences the sentence was ambiguous, reading smoothly with either a class 1 or a class 5 meaning. Most of the 20, however, were examples of the use of "a" as an elliptical construction implying "kind of," with meanings still meeting the criteria of class 5.

With the class 3 nouns there was a marked difference in article use depending on whether or not an "of" phrase followed the noun. When no "of" phrase followed, the "0" article was used for 53 per cent of the text occurrences and was acceptable for an additional 13 per cent. Use of the "0" article alone would have resulted in 34 (100 — 66) per cent unacceptable insertions. To improve upon this it is necessary to add a second article. The article "the" was used for 26

per cent of the text occurrences and would have served for an additional 14 per cent. The article "a" was used in 21 per cent of the text occurrences and would have been acceptable for an additional 10 per cent. Using a dual article, either "0/the" or "0/a" would provide an acceptable article for approximately 90 per cent of the occurrences of the class 3 nouns in the sample text not followed by an "of" phrase.

The article distribution was markedly different for the 17 per cent (75 of 426) of the class 3 occurrences that were followed by an "of" phrase. "The" was used in 65 per cent of the text occurrences and served as an acceptable article for an additional 10 per cent. Adding either "a" or "0" would bring the number of occurrences provided with an acceptable article to about 90 per cent.

Of the forty-seven occurrences of class 5 nouns, thirty-six were not followed by an "of" phrase. Of these, the "0" article was used for thirty occurrences and would have served for four more; "the" was used for six occurrences and would have served for two more. Of the eleven occurrences of class 5 nouns that were followed by an "of" phrase, the "0" article was used for six occurrences and would have served for three more; "the" was used for five occurrences and would have served for another two. The class 5 nouns included a number of nouns derived from transitive verbs, and when an "of" phrase followed it was often the case that the relation of the noun to the object of the prepositional phrase was strictly analogous to that of a transitive verb to a direct object. This is here called a "transitive relation" to the "of" phrase. Such a relation was found to obtain in most of the occurrences for which the "0" article was acceptable. Because of the small size of the sample, these figures should be interpreted as indicative only, but they suggest that a subclass might be established for the nouns of class 5 that are derived from transitive verbs, so that, when an "of" phrase follows, the dual article "the/0" will be supplied to them and "the" to the other class 5 nouns.

With occurrences of plural nouns of the sample text, the "0" article was used for approximately 78 per cent and would have been acceptable for another 13 per cent. The difference in article ratios (0:the) between plurals of class 1 and class 3 nouns was trivial. As with the singular class 1 nouns with similarly discrete referents, there appeared to be no significant difference between the article ratios relating to the presence or absence of a following "of" phrase. If the text that was analyzed does include an abnormally large number of nouns with a generic meaning (and at present we have no criteria by which to identify "normal" text), the number of plural noun occurrences requiring "the" might be found to exceed the present 10 per cent, suggesting possible future reconsideration of the dual article "0/the" for plurals.

## C. ARTICLES PROPOSED FOR INSERTION

On the basis of the foregoing analysis of intersubstitutability of articles, it is proposed to supply dual articles to singular nouns of class 1 ("the/a"), class 3 ("a/0" and "the/0"), and to those nouns of class 5 that are followed by an "of" phrase ("the/0"). A single article is proposed for all others: "the" for nouns of class 4 and the "0" article for the rest. For the 1,965 noun occurrences in the sample text, 50 per cent would receive single articles, 50 per cent dual articles, and 7 per cent of the insertions would be unacceptable.

Since it is known that the article "the" is at times required with nouns in the classes from which it has been excluded on statistical grounds, it is of interest to consider the "cost" of providing it to the nouns of these classes of the sample text: Adding "the" for all nouns of class 5 would require a trade in the sample text of 36 more dual articles in exchange for two more acceptable insertions. Adding "the" for plural nouns would require a trade of 587 dual articles in exchange for fifty more acceptable insertions.

## D. ERRORS AND REMEDIES

Three kinds of errors may be distinguished in the results of applying the above proposal to the sample text: (1) errors due to idiomatic article usage in violation of the noun classification; (2) errors due to inappropriate or imprecise coding of the noun; and (3) errors due to our present inability to select a single correct article from among the alternatives compatible with the noun classification; this failure accounts for the use of dual articles.

Correcting the first kind requires recognizing those idiomatic occurrences of nouns that require exceptional article insertion. (Of course, not all articles required within idioms violate the article coding of the noun.) Idioms are found to be of two general kinds: *(a)* those in which all words are specified—such as "of course," "for example," "in fact," "in general," "by means of," "in turn," "in favor of," "in content"— and *(b)* those in which different words (often of a semantically restricted set) may be inserted into an idiomatic frame —such as "in terms of (role)," "from (sentence) to (sentence)," "(day) after (day)," "by (telephone)," "(word) for (word)." Compilation of a list of English idioms should go hand in hand with coding nouns for article insertion, so that irregular articles can be provided on recognition of the idiom and idiomatic occurrences will not be used as test contexts in coding. For example, in the above idiom, "hand in hand," use of the "0" article is due to the idiom and should not be taken to represent normal article usage with "hand."

The second kind of errors, those due to imprecise coding, can be reduced to some extent by subdividing the present gross classes, as, for instance, by identify-

ing class 3 and 5 nouns derived from transitive verbs. Primarily, however, they are represented by the errors in article insertion for nouns of class 3, for which we are at present unable to provide mechanizable criteria for distinguishing between class 1-type and class 5- or 6-type uses. Identification of the class 1-type uses would at least permit changing the dual article to "the/a" and, so, to provide a correct article for all the non-idiomatic occurrences of this group, albeit still a dual one. Although a class 3 noun in context can usually be assigned to a more narrow article class, it is often difficult to define the determining elements, which may be elusive semantic attributes of other words or even general knowledge deriving from the universe of discourse. A clear-cut example of class determination is seen, however, in the phrases "republished in German" and "translation into Russian," where "publish in" and "translate into" require understanding the names of nationalities as language (class 5-type meaning) rather than a person (class 1-type meaning). A cumulative catalogue of such semantic indicators of the sense in which a noun is used in context will allow for a significant increase in the precision of class identification; implementation of this information will require some specifically semantic algorithms.

The third kind of error, insertion of dual articles, reflects our present inability to select a single correct article from among the alternatives allowed by the coding. What is required is to define in a mechanizable way those elements of context, implicit or explicit, that constrain article selection.

## E.   DISCUSSION OF ARTICLE DETERMINATION

Certain elements of context themselves assume the semantic function of articles. In idioms, not only is any article usually completely determined, but it may comprise an essential part of the idiom without being semantically significant per se. Those modifiers that suppress all articles with the following nouns (in general: numbers, indefinite quantifiers, demonstratives, and possessives) do so by semantically taking over the article function, as does the capitalization of proper nouns in written text.

Apart from the foregoing, it appears that the class characteristics of a noun referent, with respect to discreteness, together with its grammatical number, determine which set of articles may be used with the noun: "the" and "a" when the referent is discrete and enumerable and singular; "the" and "0" (and under certain circumstances, "a") when the referent is non-discrete, generic, or abstract and singular; "the" and "0" when it is plural.

"The" is usually, but not always, used to denote particularity. It also has a generic use, usually equivalent to use of the plural with the "0" article. This appears to be what J. Barton [2, p. 114] means: "The definite article presents the nominatum in, and with reference to, its history. It either calls upon our knowledge of the same nominatum, a knowledge derived either from previous reference, direct or indirect, in the same discourse, or from general culture; or it explicitly gives the nominatum a univocal individual specification, for example by relative clause, that is, it *provides* a history, as in 'the hat which I bought is too small.'" As Beverly Robbins indicates in an unpublished memorandum (University of Pennsylvania, Transformations and Discourse Analysis Projects, No. 38, p. 125), for "the" to be interpreted in this way it appears that "the whole sentence must be pervaded by a generalizing quality." It also appears that use of "the" with a singular noun without the expected contextual corroboration of particularity tends to *confer* a generic meaning to "the." Since, however, this is precisely the situation where the mechanical indication would be for an indefinite article, no way is seen to make use of this English pattern in machine translation when English is the target language. In fact, there seems to be no way to prescribe use of an indefinite article except from lack of indications for "the," since the indefinite article implies knowledge about the existence and rightness of the rest of the class which is independent of context.

Any article, "the," "a," or "0," may be either determined by context or used in a semantically independent way, carrying information not duplicated elsewhere in the context. The likelihood that the article choice is constrained varies with the kind of indicative elements present. As noted above, contextual evidence for "a" with class 1-type nouns, or the "0" article with class 5-type and plural nouns, is primarily negative— that is, absence of indications for "the." The presence of an "of" phrase following a noun with a class 5-type meaning that is not derived from a transitive verb is a fairly reliable indicator that "the" is required. (Restrictive clauses following nouns with class 5-type meanings would be also if appropriate English punctuation were available to the machine-translation program; unfortunately, it is not.) However, an "of" phrase, or even a restrictive clause, following nouns with class 1-type meanings and plurals is only weak presumptive evidence for "the," although sometimes it appears that context lowers the threshold for unique identification, allowing a phrase to govern selection of "the" when it would not necessarily do so if the sentence were removed from context. To deal with the semantically independent occurrences of articles it appears necessary either to retain dual articles where a single article cannot be specified, since the "0" article that results from non-insertion can be as eloquent as the explicit articles, or to follow the patterns observed to occur with highest frequency on statistical grounds alone.

In the majority of cases, however, there *is* a seman-

tic determinancy imposed by the nature of the noun referent and by context which must (redundantly) be expressed by an article in idiomatic English. The contextual determinancy may either result from delimiting the sense in which a multiple-meaning noun is used, thereby establishing discreteness or non-discreteness (i.e., the class-type characteristics) or may result from the presence of information in the light of which particularity or non-particularity can be deduced. When particularity is implied by context, thereby requiring insertion of "the," the relevant context is generally found in:

1. Certain preceding modifiers of the noun (see below, "Some Specific Rules for Article Insertion") including mainly words that have reference to quantity or specificity.

2. Certain syntactically linked modifying constructions within the sentence:

   a) Modifying phrases that follow the noun, be they participial, prepositional, or adjectival, *if* they answer to the question "which one?" rather than "what kind?"

   b) Restrictive clauses following the noun, *if* they contain identifying information.

3. Semantic context, which may be outside the sentence:

   a) Any unambiguous reference within the discourse, explicit or implicit, to the referent of the noun (usually prior to the noun occurrence, but not always).

   b) Semantic implications inherent in the setting and subject matter of the discourse, which may demand either a particularizing or a generic "the."

General criteria amenable to machine processing have not yet been formulated to distinguish either the adverbial phrase (which is irrelevant to article selection) from the adjectival one (which might be), or, in the absence of proper English punctuation, an irrelevant non-restrictive clause from a possibly relevant restrictive one. However, it is relatively easy to define and apply rules that depend on the presence of mechanically identifiable and enumerable contextual elements. A preliminary list follows.

### Some Specific Rules for Article Insertion

1. Suppress article insertion when a noun is preceded by:

   a) A possessive modifier (the possessive form of either a pronoun or a noun);
   b) A demonstrative modifier ("this," "that," "these," "those");
   c) An interrogative "which?" "what?" "whose?"

2. Suppress article insertion when a noun is preceded by: "each," "every," "any," "some," "no."

3. Suppress article insertion when a noun is preceded by the following *used as adjectives:* "much," "most," "more" (except in the idiom of two comparatives: "the——er,

the——er"), "less" (except in the idiom of two comparatives: "the ——er, the——er").

4. Insert no article after a hyphen in a hyphenated word.

5. Use "the" with a superlative, which may be a pronoun such as "the best," "the most," "the highest," etc., or a noun with a superlative modifier. The article should precede a preceding adverbial, if one is present. (There is a figurative use of the superlative, as in "a most careful computation," that is not expected to be required for machine translation in which English is the target language.)

6. Use "the" before the following: "same," "very" (used as an adjective), "only," "next" (except use "the/0" in adverbial expressions of time).

7. Use "the" with a plural noun that occurs in an "of" phrase following any of the following: "one," "each," "another," "anyone," "anything," "any," "many," "few," "several," "part," "the rest," "some," "most," "all," (any number).

8. When "such" is used as a modifier, use the following articles after "such": "a" with class 1 and class 4 nouns, "0" with class 5 nouns and all plurals, "a/0" with class 3 and class 6 nouns.

9. The modifier "one" substitutes for the article "a" but may be used in addition to the article "the." Hence the article "the/0" should be supplied to singular nouns (except those of class 6).

Information outside the sentence demanding use of "the" includes explicit and implicit reference to the noun referent. This accounts for a great many uses of "the" with class 1-type nouns and plurals in running text. The reference need not be to an identical word form or stem; it need not even correspond in gender and number as an antecedent does to a pronoun. The reference may be purely semantic, implicit rather than explicit, and comparable only in terms of abstractions. To find such reference mechanically will require inputting some representation of the semantic attributes upon which the identity is based and probably can never be done exhaustively. The task of identifying the significant ones has barely been started.

We are now able, however, to analyze why a following "of" phrase affects article use. Of the two article functions, (1) establishing discreteness or its absence and (2) establishing particularity or lack thereof, an "of" phrase affects the second. It often, but not always, confers particularity upon the referent of the noun that it follows.

With class 1-type meanings, we find that the required article can carry the full burden of establishing particularity or non-particularity, independent of any modifiers preceding or following the noun. This is true whether the noun is coded as class 1 or is coded as class 3 and used with a class 1-type meaning. For such occurrences, the presence or absence of a following "of" phrase generally does not affect the article. This

can be demonstrated by dropping or inserting an "of" phrase following class 1-type occurrences and noting that there is no concurrent need to change the article.

For class 5-type nouns, a following "of" phrase usually serves to partition the generic referent of the noun it follows, thereby particularizing it and imposing the requirement of "the," as in the phrases "the fidelity of the translation," "the grief of the mourner," "the accuracy of the calculation," "the language of computers," etc. This situation is indicated if the meaning is not violated when the object of "of" is made possessive ('s) and placed before the noun in question, as "the translation's fidelity," "the mourner's grief," etc. However, if this transformation cannot be made, as in the phrases "sand of the desert," "scrap of all kinds," "shortness of breath," etc., no conclusion can be drawn as to which article ("the" or "0") is appropriate. Hence, to the extent that such meanings are also expressed in other languages by a genitive phrase, an article prescription for "the" may be incorrect.

Further, a following "of" phrase fails to be a reliable indicator for "the" when it functions, not to partition or particularize the noun it follows, but to complement it in the manner of a direct object to a transitive verb, as in the phrases "control of the machine," "direction of the play," "transmission of the information," "translation of the article," etc. In these latter instances "the" and "0" are usually substitutable, and "0" seems often to be preferred. The distinction can be seen clearly in the following two sentences: "Admiration of the man inspired the boy." "The admiration of the man inspired the boy." Use of the "0" article causes "of the man" to be understood as object of the transitive verb "to admire," and it is the boy's own admiration that is said to have inspired him. Use of "the" causes "of the man" to be understood as partitioning the generic noun "admiration," and it is the man's admiration that is said to have inspired the boy. It appears that, for those nouns that allow it, that is, generally those derived from transitive verbs, the transitive kind of relation to a following "of" phrase tends to be more frequent, thereby justifying a semantic partitioning of the nouns of class 5. How frequently "the" is required with this group of class 5 nouns has not yet been investigated over a sufficiently large amount of text to make firm generalizations, but it appears that the "0" article is used more frequently and, further, is often substitutable for "the."

A number of such semantically defined subgroups are expected to emerge for each article class on further investigation.

## F.   CONCLUSIONS

In order to determine how further improvement can be achieved, both in terms of fewer unacceptable insertions and in terms of fewer dual articles, we have inquired into the semantic role of articles and the kind of linguistic elements that affect their use. This work has indicated that a certain amount of further refinement in the article-insertion program can be achieved by relatively straightforward and simple techniques, such as: (1) cataloguing English idioms so as to insert correct articles and to exclude idiomatic usage from consideration in coding nouns; (2) excluding from consideration in coding, for either general or subjected-restricted text, meanings that occur too rarely to warrant recognition (i.e., excluding statistically trivial "counterexamples"); (3) extending the catalogue of special modifiers and specific constructions that either preclude any article at all or make a given one mandatory.

Further progress, however, will require dealing with articles as a semantic problem—in terms of semantic attributes and semantic relations. Our work has indicated that whether or not the referent of a noun is discrete and enumerable determines its article-class assignment and constitutes the semantic datum upon which other rules for selection of article must operate. The definite article may be required by syntactically linked context within the sentence, by greater semantic context outside the sentence, or it may introduce new information. Those elements within the sentence that cause "the" to be required are phrases or clauses that contain identifying information (designating which one or which particular part as opposed to designating what kind). Beyond the sentence boundary, the existence of any unambiguous semantic antecedent of the noun usually dictates use of "the."

Hence fundamental improvement in article insertion for machine translation will depend on progress in the following areas: (4) cataloging those semantic relations, mainly between syntactically linked elements in the sentence, that restrict a multiple-meaning noun to only one article class; for example, when "translation" is the object of "read," the only appropriate meaning of "translation" is some sort of document; the meaning of "translation" as process is excluded; (5) subdividing the article classes that have been defined, taking into account those semantic characteristics that may affect article selection under restricted conditions; for example, nouns derived from transitive verbs are found usually to stand in a different semantic relation to a following "of" phrase than other nouns in the same class and to require different article treatment in this context; (6) determining under what conditions different kinds of modifying elements contain identifying information; the present study has indicated that the significant sentence elements are restrictive clauses, modifying phrases of various kinds, and a limited number or preceding adjectives and that they affect nouns of the different classes very differently; (7) finding ways to discover prior reference to the referent of a noun—that is, to identify semantic antecedents of nouns in the

discourse; this is relevant because often within the context of a single sentence whether the modifier is identifying or not is *specified by* the article, while with respect to the larger context the article itself may be determined.

### III. Evaluation of Automatic Article Insertion in Machine-Translation Output

The pattern of article insertion recommended in Section II was implemented as part of the Bunker-Ramo machine-translation program and tested on a Russian translation (from the original in English) of one of the articles of the sample text (Fig. 1). The purpose was to observe the interaction between the article-insertion routine and the rest of the machine-translation program.

#### A. RESULTS

Of the 480 noun occurrences, 91 per cent were supplied with an acceptable article, at a cost of providing a dual article to one-third of them. Seventy-one per cent of the total were supplied with articles in accordance with the noun-coding and recommended article-insertion pattern. For 27 per cent of the total the article treatment was determined in accordance with other criteria, which take precedence in the machine-translation program over the article-insertion routine based on noun-coding. Two per cent of the noun occurrences were incorrectly handled by the syntax program.

Of the 341 noun occurrences provided with articles by the article-insertion routine, 29 per cent were supplied with *all* the articles allowed by the noun-coding, with only one unacceptable insertion. For the remaining 71 per cent, one of the allowed articles was omitted, at a cost of one unacceptable insertion out of seven for this part of the group.

The 130 noun occurrences for which article treatment was handled in accordance with other criteria included the following cases: (1) nouns occurring with any of the specified list of preceding modifiers (66 occurrences), (2) nouns occurring in titles or headings of three Russian words or fewer (11 occurrences), (3) nouns flagged to bypass the article-insertion routine, since they were provided with invariant articles in the machine-translation dictionary (15 occurrences), (4) nouns occurring in idioms (36 occurrences), (5) nouns that are capitalized and that are not at the beginning of a sentence (1 occurrence), (6) nouns that are inclosed by quotation marks, parentheses, or preceded by a hyphen (1 occurrence). Application of these criteria resulted in three unacceptable insertions.

The remaining nine noun occurrences, or 2 per cent of the total, were handled inadequately by the syntax program, being Russian forms ambiguous as to whether singular or plural which were translated with an English singular form but given the "0" article appropriate

to a plural form. By chance, for one of these occurrences the "0" article was acceptable. Not included in this tally are five occurrences of two nouns that failed to be coded and two noun occurrences in passages so inadequately handled by the machine-translation program that an appropriate article could not be determined.

#### B. ANALYSIS OF ERRORS

Of seventy-six occurrences of class 1 nouns, the single unacceptable article occurred in the frame of an English idiom in the phrase "definite and unique in its *kind of (0) advantage*." The article "the/a" had been supplied. The obvious remedy requires recognition of the idiom and programing to suppress the article of the noun following "kind of."

Of the seventy-six occurrences of class 3 nouns, thirteen constituted article errors: eight occurrences out of fifty-one without a following "of" phrase were supplied with "a/0" but required "the"; five out of the twenty-five that were followed by an "of" phrase were supplied with "the/0" but required "a." The nine words involved were: "language," "order," "communication," "material," "mechanism," "translation," "study," "meeting," and "velocity." A more narrow article code does not seem advisable for any of these nouns, with the possible exception of "mechanism," which is probably used without an article only in philosophic discourse.

Of the twenty-eight occurrences of a class 5 noun with no "of" phrase following, a single error occurred in the phrase "that the address actually received or understood *(the) information* sent him." The "0" article was supplied, but "the" was required by prior reference to the information.

The 139 occurrences of plural nouns were all supplied with only the "0" article. Nineteen were in error, requiring "the."

The one occurrence of a class 4 noun, the thirteen of class 5 nouns that were followed by an "of" phrase, and the eight of class 6 nouns were all supplied with all the articles for which they were coded and included no errors.

The 130 occurrences for which the article was determined by other criteria included three errors. One was due to including "such" in the list of modifiers that always cause articles to be suppressed. The remedy is to provide for inserting "a" after "such" with class 1 nouns, "a/0" with class 3 nouns, and the "0" article with class 5 and plural nouns. The rule to omit any article before a capitalized noun in the middle of a sentence led to one error: "Accuracy was estimated by a judge expert who used the criteria of *(the) State Department . . .*" Probably most such cases can be handled as idioms or by recognizing capitalization as a variable in noun-coding. Although it caused no errors

KNOWING BOTH LANGUAGES, WERE AT OUR DISPOSAL AS EXPERIMENTAL OBJECTS. A/O MATERIAL, WHICH USED
  TONGUES                    IN QUALITY    ITEMS              WHOM   ENJOYED
                                           OBJECTIVES                UTILIZED

*id*

WRITTEN AND ORAL TRANSLATORS, WAS A/O POLITICAL MAINTENANCE.. PUBLISHED REPORTS OF MEETINGS OF
WRITING                          CONTENT                                        RECORDS
                                                                                MINUTES

*id*        *dict*

THE SECURITY COUNCIL OF THE UNITED NATIONS . AT THE TIME OF EVERY EXPERIMENT WE NOTED THE/O PRODUCTIVITY OF
                                                            EACH                    MARKED THE/O OUTPUT
                                                                                          THE/O CAPACITY

*id*

LINGUISTS FROM THE POINT OF VIEW OF THE/O ACCURACY AND OF VELOCITIES, EXPRESSED BY THE/A NUMBER OF WORDS IN
                                                          SPEEDS     INDICATED                           TO
                                                                                                        AT
                                                                                                        ON

(caps)

THE/A MINUTE. ACCURACY WAS ESTIMATED BY THE/A JUDGE EXPERT, WHICH USED CRITERIA OF STATE DEPARTMENT, USED AT
                                                            WHO                                                FOR

THE/O SELECTION OF PROFESSIONAL LINGUISTS.
THE/O SEPARATION

Title: RESULTS. WE STUDIED TWO PRINCIPAL LINGUISTIC METHODS AND SEVERAL THEIR VARIANTS. FIRST
                INVESTIGATED  BASIC                                                     VERSIONS
                LEARNED

(A) (Rpl.)

EXPERIMENTS WERE CONNECTED WITH THE/A UNIVERSAL METHOD OF A/O TOTAL TRANSLATION. GROUP OF TRANSLATORS, SOME
                                                              CURRENT COMPLETE TRANSFERENCE

Fig. 1.—Sample page from machine-translation printout incorporating article-insertion routine. *Check*, acceptable article; *circled word*, a required article not provided; *id*, article provided as part of an idiom; *dict*, article provided as part of the English equivalent in the mechanical-translation dictionary; *(Rpl)*, Russian plural *(syntax error)*; *vertical arrow*, belongs with second English equivalent (printed where a "0" article for the first equivalent should be); *double underlining*, a modifier that suppresses all articles with the following noun.

in this text, it may be noted here that the rule to omit articles with nouns in short titles will certainly lead to incorrect insertions at times. The rule to omit articles with a noun that is preceded by a hyphen appears to be on much firmer ground. The rule to omit any article for nouns occurring in quotation marks resulted in an error in the sentence "the condition of '*(the)* inverse linguistic *problem*' had a tendency to slow down the work of the translators." This rule can only be justified on statistical grounds, and it appears to be of doubtful validity.

The additional rules proposed in "Some Specific Rules for Article Insertion" (above) were not programed.

However, in this brief text they would have found little application. The one error with "such" has been discussed above. Recognition of a superlative modifier would have eliminated one error with a plural noun: "Participants of the conferences preferred to negotiate with the help of *(the) most impersonal means* (pl) of communication." The errors resulting from supplying to an English singular form the article appropriate to a plural are not, strictly speaking, article-insertion errors. They do, however, emphasize the dependence of the article-insertion routine upon correct syntactic analysis.

### References

1. Martins, G. R. "Preliminary Report on the Insertion of English Articles in Russian-English MT Output," *Mechanical Translation,* Vol. 8 (1964).
2. Barton, J. "The Application of the Article in English," *Proceedings of 1961 International Conference on Machine Translation of Languages and Applied Language Analysis,* Vol. 1. London: Her Majesty's Stationery Office, 1962.