[Mechanical Translation, vol. 8, No. 1, August 1964]

Preliminary Report on the Insertion of English Articles in Russian-English MT Output*

by G. R. Martins, Technical Staff, Bunker-Ramo Corporation

Research on a non-statistical scheme for the insertion of English articles in machine-translated Russian is described. Ideal article insertion as a goal is challenged as unreasonable. Classification of English nouns, simple syntactic criteria, and multiple printout are the scheme's main features.

One of the most discussed problems in the automatic translation of Russian documents into English is the insertion of English articles in the output. Approaches to the solution of this problem, where it has been considered at all, are as varied as the basic MT programs in use by the different teams engaged in this work. Most projects, however, either use statistical criteria in the determination of English articles to the exclusion of all other considerations, or use a combined syntactico-statistical method; the aim of all such routines is the selection of one and only one of the four articles *(a, an, the, Ø).* None of the solutions presented to date in the literature is entirely satisfactory.

Two kinds of ambiguity present themselves as obstacles to the successful determination of English articles in automatically translated Russian. The first derives from the structure of the Russian language, in that it does not employ any simple elements isomorphic with English articles as adjuncts to nominal phrases—there are no elements in Russian text which may be correlated strongly with the English articles. This kind of ambiguity is not always formally resolvable since it often raises the particular question: "What did the author *mean* in this instance?" In such instances, even with his immense reservoir of repertorial and contextual clues, the human translator can only make an educated guess, and the machine, with its drastically limited set of potential determiners, cannot do better.

Rut another kind of ambiguity arises from the side of the English output itself. Situations are frequently encountered in which various articles may be inserted without doing violence to the text, and occasionally without altering in any simply statable way the intuitive meaning of the passage. In: "He is working on —— analysis of English verbs." we may read *an, the,* or *Ø,* with appropriate intonations, and get reasonable English sentences which differ in meaning, if at all in any systematic way, very slightly indeed. The question: "What is the preferred English article?" in these situations is not easily answered, and it does not seem a reasonable hope to look for a single arbitrary choice which will work in every case.

Here we are faced with two kinds of overlapping ambiguity, neither of which is easily resolved even by the human translator, and which appear to be well beyond the reach of MT machines as presently programmed.

These considerations have led me to the conclusion, surprising perhaps to some, that it is both *impossible and undesirable* to attempt the automatic determination of a single English article appropriate to the occurrence of every nominal encountered in the output text. Which is to say that we should be prepared to do without articles altogether, or to accept alternative articles in the final printed translation. The former solution, presently in use by some teams, is not quite so harmless as it appears, for the reason that Ø is as legitimate an English article as arc *the, a,* and *an,* to my way of thinking. The decision (or pseudo-decision) to do without articles altogether, then, amounts to a decision to select everywhere the article Ø , and this is scarcely more defensible than to select everywhere *the* (which is statistically much more common).

The decision to print out alternative articles in some instances is tantamount to passing on a portion of the translation function to the reader, of course. While this hardly fulfills the idealists' goal for MT, it is not an indefensible solution; the same default of function can be imputed to every MT program which permits multiple printout as a solution to very complex problems of polysemy—and this includes every existing program. And, so long as (a) we do not simply print out all four possible articles in every case, and (b) we do not fail to include among the output alternatives a/the "correct" article, we have made a net gain in quality of translation. What is more, the task of final article selection might, in most cases, better be assigned to the reader, knowledgeable of the field of discourse and possibly even familiar with the stylistic peculiarities of the author, than to the machine.

This point of view not only enables us to proceed in spite of the ambiguities mentioned above, it gives us at the same time one of the distinctive characteristics (multiple printout) of the system we have been looking for as a solution to the article problem.

It may legitimately be asked at this point whether the net translation quality gain obtained even from the best of multiple-article-printout schemes justifies the research and programming effort required for its implementation. From the point of view of a produc-

* This research is being carried out under the sponsorship of the National Science Foundation.

tion MT organization, this question is meaningful only in terms of the incrementing of consumer appeal of the product, and it would be difficult to answer without research in that very area. From the point of view of an MT research group, the implementation of such an article insertion program as that discussed here is justified as a test of the program's inherent merits and also as a means of facilitating research into the question of consumer reaction to it.

With these thoughts in mind, a close examination of several texts, in English, was undertaken to determine something about the patterns of occurrence of the articles. Some simple contextual criteria were sought which would enable us accurately to predict the human translator's selection of an article; at this point, our attention focused on English texts translated from the Russian, and the matching Russian texts, rather than on random English texts. Decision criteria were sought in both languages in the hope that this would improve the odds on our success.

Early in the study one criterion of great promise came to light. For each English noun token in the text we asked the question: "Is its Russian equivalent, in the matching Russian text, followed by a syntactically linked genitive block?" More obvious, of course, but of great importance, was another criterion: "Is the English noun token singular or plural?" To test the significance and power of these two criteria, and to gauge the strength of additional criteria that might be necessary, the following test was devised.

A machine-translated corpus, taken from *Pravda,* was treated in the following way: (a) the corpus was divided roughly into two halves, (b) all English noun tokens in the final half were marked to indicate whether or not the Russian equivalent was followed by a linked genitive block, (c) all articles already present in the English were deleted, (d) appropriate article tokens were then inserted in the English by hand, with multiple entries being made where no clear decision could be made on the basis of individual sentence content alone, (e) each noun from the text was then listed along with indications of the article patterns occurring with it (note that here two separate entries in the tabulation were made for a noun if it had occurred in the text both with and again without a following genitive block behind its Russian equivalent), and (f) the tabulation was examined for possible clues to additional criteria.

Encouragingly, it turned out that the English nouns could be grouped into five classes according to the pattern of article occurrence indicated for them in the tabulation. This was regarded as encouraging because, first of all, three of the classes were quite small compared to the others, and secondly, each class seemed to have its own intuitive internal homogeneity.

The first half of the corpus then had its articles deleted throughout, and, for each noun in the tabulation, articles were inserted with reference only to the criteria just developed. In no case was an unacceptable result obtained from this brief test.

After this, the nouns occurring in the first half of the corpus but not in the second (and therefore not tabulated) were listed and each was classified intuitively as a member of one of the five article-pattern classes. Once again the first half of the corpus was tested, and again no unacceptable results were obtained. It is worth noting here that noun tokens occurring in special word combinations or idiomatic expressions were not taken into consideration; no particular problems are presented by such occurrences since our present MT program takes such constructions into account already for other purposes.

Other syntactic criteria, of the most obvious kind, were taken into account during these tests; these do not seem to be of such great interest as to warrant discussion at length. Typical of these criteria is: 0 with all nouns preceded by a possessive pronoun, or by a demonstrative, or by the interrogative "WHICH" or "WHAT", or by "EACH" or "EVERY" or "ANY" or "SOME". Another example is: *THE* before a superlative modifier (and before a preceding adverbial, if such is present) *.

I am pleased with the results of these early tests of the article determination procedure for several reasons. First of all, it seems reasonable to think that a successful article determination program would be based upon a classification of English nouns and upon certain rather simple syntactic criteria; this is the approach hinted at by the Milan MT team, although their report is distressingly vague and little more can be got from it than the fact that they are thinking in terms of eight noun classes, not five.[1]

The intuitively satisfying homogeneity of the contents of each noun class leads me to suspect that such classification as we are undertaking could have some relevance outside the restricted domain of MT. A related consideration is the apparent success of attempts to classify nouns intuitively; this not only raises certain mildly interesting questions about the grammar of English, but it greatly enhances the feasibility of carrying out such classification in extenso.

To make clearer some details of the scheme, I will give here a set of noun-classification rules put together earlier in our study to serve as a research tool. The following rules are suggestive rather than strictly prescriptive in nature. It is hoped that rules of this kind will enable linguistically unsophisticated personnel to carry out successful classification operations on the membership of large noun lists without time-consuming context consultation and/or revisions based upon hindsight. A small burden is deliberately placed upon the worker's imagination, and it is presumed that the worker is a native speaker of English. These restrictions are felt to be justifiable for two reasons: (a) we thus avoid the premature elaboration of very complex

* The obvious exceptions to a rule of this kind for mathematics texts are now under study.

rules, and (b) the worker's imaginative burden diminishes rapidly with experience in this kind of coding operation.

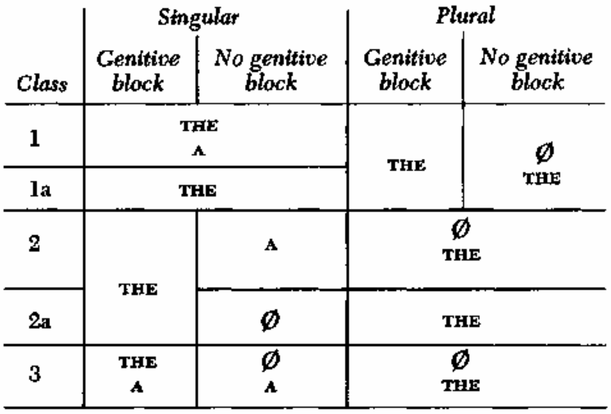
The rules take the form of simple questions, answerable with either "YES" or "NO". Coding indications depend upon these answers.

1. Can the noun, in the singular, begin a sentence of the type: "—— is necessary." etc.?
YES: See rule 2
NO: See rule 3
2. Can the noun, in the singular, ever take the article "A/AN"?
YES: Class 3
NO: Class 2a
3. Does this noun, in the singular, *always* require "THE"?
YES: Class la
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 diagram in the next column, with an accompanying explanation, shows the relationships between the noun classes thus established and the article selection routines.

Reference

1. J. Barton. The Application of the Article in English. *Proceedings of the 1961 International Conference on Machine Translation of Languages and Applied Language Analysis (Teddington)*, Vol. I, Her Majesty's Stationery Office, London, 1962, pp. 111-121.



	Singular		*Plural*	
Class	*Genitive block*	*No genitive block*	*Genitive block*	*No genitive block*
1	THE A		THE	Ø THE
1a	THE			
2	THE	A	Ø THE	
2a		Ø	THE	
3	THE A	Ø A	Ø THE	

Explanation:

English nouns are classed by membership in one of the five classes listed in the leftmost vertical column of the diagram; a very small number of special nouns are not so classified, but are covered by individual rules (e.g., "mankind"; NO ARTICLE). The categories "Singular" and "Plural" refer to the noun token itself. The indication "gen. block" means "noun token is followed (in the Russian) by a linked genitive block"; "no gen. block" is the negation of "gen. block". The listing of two forms in a section of the diagram means that both are to be printed out as alternative readings. Where 0 occurs alone, nothing is to be printed; where it occurs as an alternative reading, an indication of the alternative article-less reading is to be printed along with the given article.

Unquestionably, the simplicity of the single major syntactic criterion (relating to following genitive blocks) will have to be weakened in favor of more sophisticated criteria; but it is interesting how much of the problem can be managed with no more than this. A program is now in preparation which will permit large-scale testing of these proposals on a variety of corpora automatically; we are looking forward eagerly to these results of those tests.