An evaluation of the usefulness of machine translations produced at the National Physical Laboratory, Teddington, with a summary of the translation methods.

J. McDaniel, W.L. Price, A.J.M. Szanser, D.M. Yates

Introduction

The machine translation project at the National Physical Laboratory (NPL) has been terminated. It has always had as its prime aim a demonstration of the practicability of translation by computer of Russian scientific texts into English. In order to test how far this aim has been fulfilled and further, to provide evidence to guide a potential agency interested in giving a machine translation service, we have carried out an evaluation experiment on our translations, the conditions of which as far as possible emulated those of a translations service.

The results of this experiment are presented in this paper, together with a summary of the translation methods used. The paper as a whole will thus give an independent presentation of "what methods produced what results". For a comprehensive account of the NPL translation techniques, see reference 1.

Evaluation of Translations

We have been concerned with the translation of scientific Russian texts only. In considering how we might evaluate the results of our work, the context of use of scientific translations imposed two main constraints. Thus, firstly, in the vast majority of cases we would expect readers of translations to be themselves experts in the subject matter of the material translated, i.e. they would be reading the translations because these reflect their main professional responsibilities. We may then expect that the inherent background knowledge of such readers will ensure a high impetus to their comprehension of translations and help them through syntactic awkwardnesses and multiple-meaning choices. We would also expect that only a small percentage of these readers would have any competence in Russian. Secondly, the items of translation being read by the above typical readers will normally be whole information units (journal article, chapter of book, abstract, review, &c.), and they will have the freedom to ignore unimportant sections of such units and to use sentence or paragraph context (or even remoter references) to help elucidate obscure sections. More specifically, a particular sentence may be poorly translated, but because the reader can see that this is not an important sentence or because the context of (hopefully, better-translated) neighbouring sentences clarifies its meaning, that sentence may not affect at all an adequate comprehension of the whole.

Both these constraints are reflected in our evaluation experiment. We ensured that our evaluators were expert in the

field of the material they were evaluating, and also that they commented on the adequacy of an information unit as a whole, not on individual sentences.

We have included in this paper (FIG. 2(A)) a short passage from one of the evaluated translations, as the full translation is inappropriate for this printed version. However, the full translation will be available for inspection at the presentation of the paper, or the full translation of another paper can be examined in reference 1.

The evaluation experiment

In order to fulfil the first constraint above, we invited practising scientists to send in Russian papers, reflecting their professional speciality and preferably in the fields of general physics, electronics, or electrical engineering. Some papers resulted from direct invitation, others resulted from an open invitation published in our house journal, "NPL Quarterly". We undertook to send them the machine translations of their papers in return for their comments on how useful the results were. We also obtained second opinions from other specialists in the subjects concerned.

These evaluators were therefore as far as possible typical of the "customers" of a production MT service; in particular they had a personal interest in the subject matter and usually little if any knowledge of Russian.

In all 44 papers were received in response to our invitation; of these 28 were translated in full[1]. Seven of these were disregarded for various reasons[2], and the remaining 21 were included in the evaluation.

38 comments were received on 19 of these 21 papers. Of these two were rejected for vagueness, and three brief comments from one group were treated as one, so in all the experiment produced 34 comments on 19 papers.

---

[1]The other 16 are accounted for as follows: 1 was on a remote subject; 2 were deferred since we had already translated three papers for the same 'customer'; 4 were withdrawn; 3 were translated only in part; and 6 were not reached by the date our computer was scrapped.

[2]3 were on inappropriate subjects; 2 were translated only by an earlier version of the programs; and 2 were translated too late for inclusion.

We had decided to give our evaluators a free hand in discussing the usefulness to them of translations of this quality. This meant that a scale had to be devised by which their comments could then be graded by us. A scale recently published in the U.S.A. (reference 2) was considered but not adopted since we felt that for our purposes more space should be given to the middle range of the scale. The following wording was adopted:

8 Fully adequate. Meaning immediately clear, even though not always conventionally expressed.

7

6 Mostly very good. A few sentences obscure, so that something essential may be lost, but normally clear enough.

5

4 Fair. Takes a good deal of time to extract meaning and even then there is no great confidence in it, resulting in a partial understanding.

3

2 Poor. Could only be useful to someone prepared to struggle hard, and even he would often be disappointed.

1

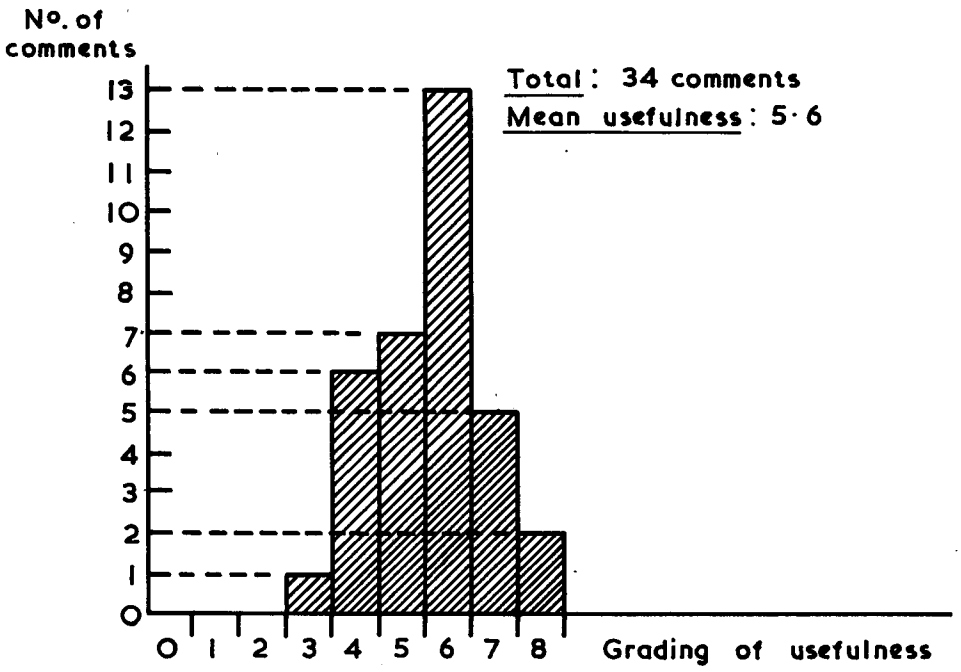
0 Useless. Although some semblance of meaning may appear occasionally, it would never be worth the trouble of finding it.

The wording of this scale is not derived on any scientific basis, but it has proved useful in practice, since when four of us came to grade the comments by it independently, there was a good agreement between our markings. Our four individual ratings for each comment were reduced to a single rating (normally the mean) after discussion. The range of scores is shown in FIG. 1; the mean score is 5.6.

The spread is no doubt due to a real variation in the quality of the translations combined with the prejudices and degrees of patience of the evaluators. The lowest scores thus come from impatient professional translators dealing with a poorer-than-average text, while the highest ones are perhaps over-enthusiastic supporters dealing with a better-than-average text.

The consensus though, is that there is a real demand for translations of this quality, and this result provides, we feel, ample justification for mounting a broader evaluation exercise, over a wider range of potential readers of such translations, to strengthen, if possible, this verdict and make it possible to

FIG. 1 Assessment of usefulness of N.P.L. MT output.



decide on the viability of a production machine translation service based on our system.

Evaluators' criticisms

Apart from the opinions as to the general usefulness of translation, evaluators' comments contained many particular points of criticism which deserve discussion. We are able to comment ourselves on some of these points from the position of having done considerable development work, just short of full implementation, on techniques designed to overcome the particular translation faults. Full details of this further work are given in reference 1, and specific points of reference are given below.

Most of these criticisms can be classified into three groups, concerning respectively: (i) the English equivalents offered, (ii) the syntactic resolution and (iii) the word order.

A frequent criticism concerned missing or inappropriate equivalents. In addition to fully justified remarks of this kind there were also cases in which the meaning proposed, or preferred by the reader, was uncommon. Its absence from the dictionary was the result of a preferential choice having been made, a compromise between completeness and simplicity. The other alternative, including all possible equivalents, would of course drastically impair readability. The particular solution is often very difficult and can only be achieved to a satisfactory degree after long experience.

In other cases there is no obvious preference and the problem is further aggravated by the very high frequency of occurrence of the word. Here belong some special classes, for example all prepositions and some very common words such as и , а , and что . Prepositions can and should be resolved by considering them together with either the governing word or the governed complement (nominal or otherwise)[1]. (For example, увеличить.. на.., 'to increase .... by ....'). For the awkward common words specific syntactic sub-routines should be devised. In practically all cases the solution is unique (see reference 1).

Only two evaluators complained about the necessity of selection among two or three equivalents. This is a matter of preference, but it seems to us that for a bona fide reader an additional possibility of meaning (if it is not carried too far) is more an asset than a disadvantage, even if it impairs to some

---

[1] On the lines already used for the recognition of idioms, expanded to include non-adjacent words; see below in the summary of methods.

extent smooth reading[1]. Until a semantic analysis can be achieved, multiple equivalents are bound to stay in MT.

A minor point, but nevertheless worth attention, was to the effect that when multiple equivalents followed each other, the difficulty in understanding increased out of proportion. For example, случается при appears as: 'occurs in ' when the results with
actual meaning is often 'results in'. This was undoubtedly a real problem, which could perhaps be helped by using a longer space between sets of multiple equivalents in the output.

Complaints concerning un-idiomatic translations (e.g. 'period of work' instead of 'life-time') would be allayed by more work spent on our idiom list, which contained only about 540 items, whereas 1,500 would be a more realistic figure.

Complaints about inadequate syntactic analysis, leading to obscurities, ambiguities, and wrong resolutions, would have been considerably reduced by a full implementation of the syntactic routines described in reference 1. One of the minor but annoying ambiguities, which had been resolved theoretically, but only partially implemented, was that of adverb/short adjective. Order of clause components was a frequent subject of criticism; of course they can be re-arranged according to the English usage only after a complete analysis has been made.

Among other things criticized was an inadequate treatment of abbreviations and abbreviated units, some of which were covered by dictionary entries, while others were not, and this led to some misunderstandings. Obviously this again is a matter for a more complete dictionary[2]. The most difficult case is "nonce" abbreviations (we met, for instance, нейтр. for нейтронный and produced 'non-itr.', which helped no one!) Here we see no prospect of a solution.

Our "anglicizing" routine was criticized (while appreciating the general idea) for unorthodox transliteration, which made it more difficult to identify the word in a standard dictionary, if necessary[3]. A partial solution may be to exclude certain word

---

[1] Much can be said on this point. Readers, no doubt, will realise how a velvet smoothness of translation may hide many a grievous fault.

[2] With a few exceptions, however. Thus 'в' may be very troublesome, as regards the choice between the preposition and the abbreviated unit ("volt"), without a special syntactic subroutine.

[3] This criticism clearly implied some knowledge of Russian.

classes, e.g. acronymic abbreviations, which are obviously not suitable objects for the routine (they can be automatically recognized as clusters of capital letters). Also, in our prefix-recognizing routine there is an inherent danger that a "not-in-dictionary" word may have a part of the stem identical with an accepted prefix. This applies in particular to short prefixes, like не- , in the above example of нейтр.. There is no general way of dealing with such words. The best solution, in respect of both routines, seems to be, however, to include in the output both the original (in Cyrillic, if possible) and the synthetic equivalent for all "not-in-dictionary" words.

A few comments contained bouquets rather than brickbats. One evaluator commented that the translation became easier to read as he got used to the unusual 'style'; and another found an instance where a slip in the published human translation had reversed the intended meaning; our version of the passage, while not perfect by any means, was certainly not misleading in this way.

Finally, several evaluators commented that machine translations would need to show advantages in cost and speed over human translations in order for them to be attractive as well as acceptable, and these are indeed criteria that we would ourselves put forward without fear of contradiction. We have not included a study of cost and speed within this evaluation experiment, as we do not have the market data to prepare a translation service specification that we could then refer such a study to. However it is evident that our machine equivalent of the human translator i.e. input punching, machine translation and output printing (with no human post-editor) will show a clear advantage on both these points. It would be essential to fit this component, though, into an overall translation system which was specified carefully to fit the translation market.

In FIG. 2(A) is shown a facsimile of a short passage of our machine translation into English of a Russian text on electric furnaces, completely non-post-edited. The vertical lists of two or three words are to be read as alternative English correspondents for the Russian word in that position. FIG 2(B) is a facsimile of the original Russian text.

A summary of the translation methods

Text Preparation and Dictionary Look-up

The dictionary used in the NPL machine translation system was developed from an early version of the Harvard Russian-English computer dictionary. Our dictionary contains about 18,000 entries (with additional cross-reference entries) covering the fields of electronics and electrical engineering.

We chose to organize the dictionary on a stem and suffix

Metal melted into furnace(s) is possible to present in the form of continuous
in

block , but then to cut out from it elementary cube of any dimension and to
assembly and engrave size also

define its resistance.
determine

Having replaced elementary cube of melted metal by unit of electrical circuit of
node
knot

model, is possible to reveal distribution of current in it and , having
also

modelled thus all bath of furnace, is possible to recognize character of
learn

distribution of current in melted metal.

Constructional grid model represents geometrically similar volume of bath in
constructive

significantly decreased scale.

FIG. 2(A) English machine translation

FIG. 2(B) Russian original text.

распределения электрического тока ванне расплавленного металла

(андидат техн. наук, доц. А. И. ЛЕУШИН
шевский индустриальный институт им. Куйбышева

лектрического то-
ической печи, как
ольшое теоретиче-
Поэтому вопросы,
сояна [Л. 1], тре-
изучения. Особую
зание распределе-
ым металлом. Вы-
анне печи настоя-
омплексного реше-
о перемешивания
еление характера
дать более рацио-
и конструкцию
числе и размеще-
х печах, правиль-
и электромагнит-
в. В настоящей
ся применительно

авленном металле
от формы ванны,
ружения электро-
лектрический ток
характеризуется
уравнением Ла-

$=0$

еды;
кого поля.

тока через мас-
зается действием
енным точкам на
ть вектор плотно-

ектрическая цепь

ления тока в сплошных проводящих средах применяются три вида моделей: проводящие пластины, или листы, электролитические ванны и решетки, или сетки. Использование решеток, или сеток, из сопротивлений имеет бесспорное преимущество по сравнению с остальными способами, так как позволяет непосредственно исследовать распределение токов в модели.

Для осуществления подобия физических процессов объекта и модели необходим правильный выбор критериев подобия. Необходимые и достаточные условия подобия физических явлений устанавливаются третьей теоремой подобия, доказанной еще в 1930 г. М. В. Кирпичевым.

Общий критерий подобия

$$k = l\sqrt{\omega\gamma\mu},$$

где l — линейные размеры;
ω — угловая частота;
μ — магнитная проницаемость;
γ — удельная проводимость.

При равенстве ω и μ объекта и модели наиболее важными критериями подобия являются размеры и проводимость материала сетки модели.

Расплавленный в печи металл можно представить в виде сплошного блока, а затем вырезать из него элементарный куб любого размера и определить его сопротивление. Заменив элементарный куб расплавленного металла узлом электрической цепи модели, можно выявить распределение тока в нем и, смоделировав таким образом всю ванну печи, можно узнать характер распределения тока в расплавленном металле.

Конструктивно сеточная модель представляет собой геометрически подобный объем ванны в значительно уменьшенном масштабе.

Сопротивление элементарных кубиков металла имитируется сопротивлением соединительных проводов — шага ячеек модели. Шаг сетки зависит от геометрических размеров объекта и модели и, следовательно, от общего количества ячеек модели. Точность моделирования будет тем выше, чем больше число ячеек. Однако слишком большое число ячеек ухудшает условия измерения и увеличивает габариты модели и материальные затраты на нее.

basis, in which each entry contains a Russian stem together with a coded list of suffixes which can combine with the stem. This gave far fewer entries than would have been found in a full-form dictionary covering the same words. Each entry contains grammatical data and English equivalents of the Russian.

The stem and suffix organisation demanded that we create a system of splitting Russian words consistently into stem and suffix, fully described in Davies & Day, (1961). The split is made at the point determined by the maximum number of letters which together form a Russian suffix or string of suffixes. The maximum split technique sometimes causes too many letters to be treated as part of the suffix, in other words, the split is made too early in the word. Such words are provided with a cross-reference dictionary entry which directs the search to an entry in which the full information for the word is contained.

The dictionary is recorded on two reels of magnetic tape, the entries being arranged in alphabetical order. Time of consultation of the full dictionary is from 12 minutes upwards, depending on the number of entries being sought.

A text for translation is first punched on cards by an operator who recognizes Cyrillic characters, though she cannot read Russian. Symbols, punctuation marks and Cyrillic characters are represented by one card column per character. Provision is made for indicating a space to be left in the text where an equation or group of symbols occurs. These will be inserted in the translation by hand. The cards are treated as a continuous medium, card boundaries being ignored. By this means quite a long paper can be encoded on a relatively small number of punched cards.

The text, now on cards, is fed into the computer. The first computing process gives a serial number to each text word and then splits the word into stem and suffix. When all text words have been subjected to this process, they are then sorted into alphabetical order. This is essential for optimum speed of look-up in our serially organised dictionary.

The next programme in the translation sequence, the look-up programme, scans simultaneously through the dictionary and the sorted text, seeking dictionary entries corresponding to the text words. The programme allows for the occurrence of stem homographs and for the correct handling of cross-reference entries. The output of the programme (which we call, following Harvard, the augmented text) consists of the text words each with the relevant dictionary entries appended.

Having obtained a set of augmented text entries, the translation sequence then sorts these back to text order, using the text serial number originally allocated to each text word.

The result of this series of operations is a text in the original order, with dictionary entries appended to all but a few of the items. Symbols and punctuation marks do not, of course, have corresponding dictionary entries, and there may be words in the text which are not represented in the computer dictionary. The latter are given special treatment in the syntactic routines and translation output.

Provision is made in the dictionary for the representation of idioms, using a method analogous to that used in an ordinary dictionary. A "key word" is chosen in the idiom (normally the least frequently occurring word), the idiom being represented in the dictionary entry of the key word. The representation includes a list of the component words of the idiom, using which the presence of an idiomatic text word sequence can be detected before attempting any syntactic operations on the augmented text. The dictionary entry including the idiom contains the preferred English equivalent. The dictionary includes coding for 540 idioms.

Words not represented in the dictionary are given special treatment, as mentioned above. All text words which commence with one of a set of 137 Russian prefixes are looked up both with and without prefix. If the prefixed form does not occur in the dictionary, but the unprefixed form is found, then the entry for the unprefixed form is included in the augmented text, coupled with an English rendering of the Russian prefix. Despite this provision, some text words will not intersect with the dictionary. For these an attempt is made to determine part of speech, case, number, etc., by an inspection of grammatical and derivational suffixes. In the translation output the stem of the not-in-dictionary word is transliterated, aiming to anglicize as far as possible the original word. A derivational suffix is given its English equivalent in the output rendering; any prefix that was recognised is also given its English rendering.

From an augmented text produced by the foregoing procedures it would be a simple mechanical process to achieve a word-for-word "translation". We felt this was not worthwhile, as the application of relatively simple rules of grammar and syntax greatly enhance intelligibility of such a product.

Russian Analysis Algorithm

In the first place we designed and implemented a system of noun blocking and a simple predicate analysis. The results obtained were not by any means ideal, but we were encouraged to extend and refine our syntactic processes. In our first attempt the functions of Russian analysis and English synthesis were closely interwoven. As our syntactic procedures were extended to cover more features it became evident that it was essential to separate the functions of analysis and synthesis. In order to make this possible the linguistic model, described

in Yates (this conference) was developed. The model permits the analysis routines to express the Russian syntax as far as necessary and facilitates a transformation to the corresponding English sentence structure.

The analysis routines operate in a succession of passes through each sentence, defined by major punctuation mark boundaries (full stop, question mark and semi-colon).

The functions of the successive passes are as follows:

1. A preliminary pass which establishes from the augmented text the terminal element for each discrete member of the sentence. Punctuation marks are indicated in the elements for preceding or following sentence items, according to a set of formal rules.

2. A pass whose prime concern is the determination of nominal structures, i.e. nouns and words with which they are closely connected, such as adjectives or prepositions.

3. A pass which establishes links between adjacent nominal structures; the linked elements include genitive qualifiers and prepositional group qualifiers.

4. A pass which searches for potential coordinating conjunctions and examines the sentence elements or structures separated by such conjunctions, setting up coordinate groups where appropriate.

5. A pass which creates simple predicate structures, searching for words with a verb role and then locating adjacent sentence elements or structures acting as verb adjuncts.

6. A pass whose function is to examine the role of some of the more "difficult" words such as the verb **быть** and its inflected forms, and the personal/possessive pronouns **его**, **ее** and **их**.

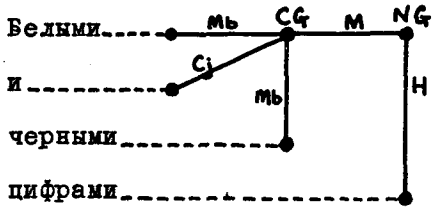
A full description of these analysis routines is given in reference 1. In the present paper we shall take a Russian sentence and note the effect of each analysis pass on it.

The Russian sentence reads:

Белыми и черными цифрами и стрелками показаны места записи границ.

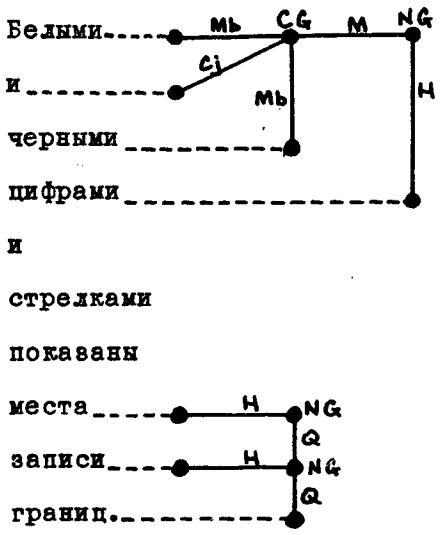
The first analysis pass is not of particular interest in the present context. Suffice it to say that a system of reference addresses is set up which permits the scanning of the sentence whilst its structure is in an incomplete state.

The state of the sentence diagram after the second pass has been completed is:-



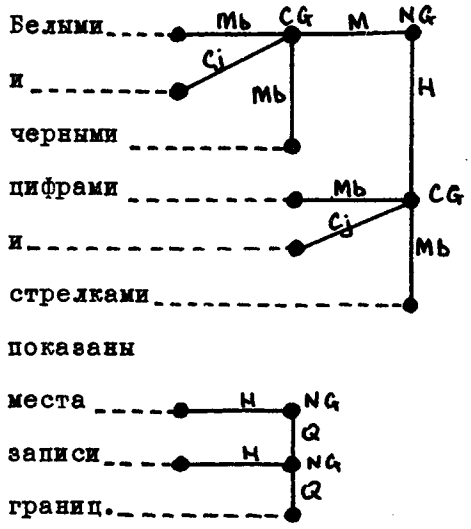
One noun group has been formed, of which the modifier is a coordinate group of adjectives. Each adjective is marked as a member in the coordinate group, which itself assumes the properties of an adjective.

The third analysis pass has the function of creating genitive and prepositional links. Only the former are concerned in our sample sentence:-



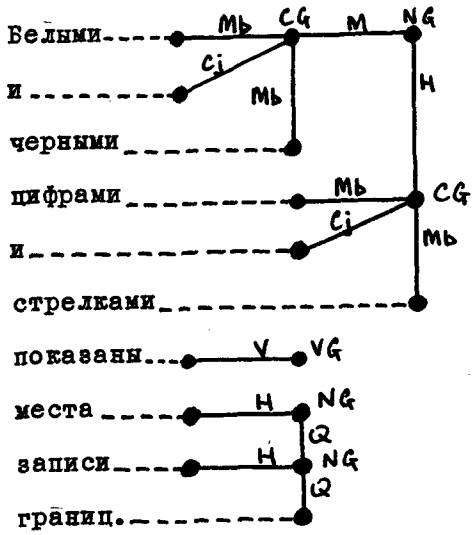
Were there any prepositional groups following nouns, then the prepositional groups would also be linked in as qualifiers.

The second analysis pass ignored all conjunctions which did not occur explicitly within simple noun groups (i.e. groups with a single noun as head). The fourth pass, however, seeks to join to existing noun groups any other nouns linked by coordinating conjunctions:-



In addition the pass groups together in coordinate groups any similar words joined by coordinating conjunctions, whatever their part of speech. Intervening punctuation prevents the formation of coordinate groups. The coordinate group, when formed, is given the grammatical significance of its component parts.

The fifth analysis pass has little effect on our sample sentence. The plural, short form participle is the single "verb" member of its verb group:-



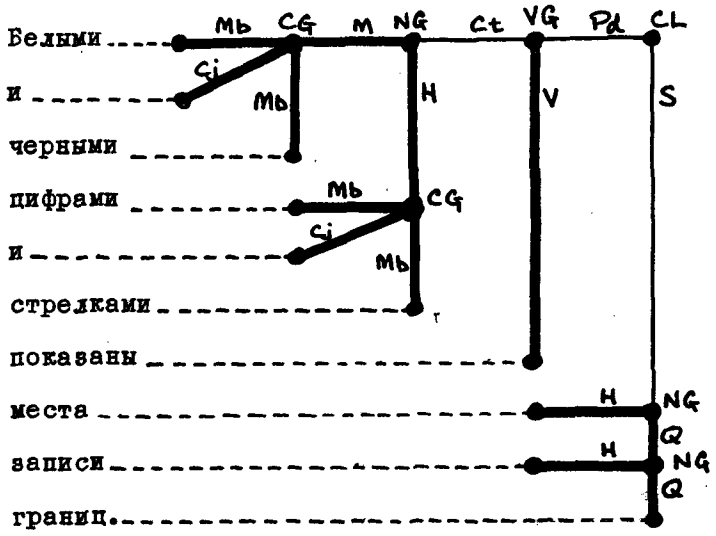
Were there adjacent adverbs or prepositional groups, these would be included in the verb group with the role of adjunct. The fifth pass also has provision for negative and conditional predicate structures.

The function of the sixth pass is to try and resolve the roles of certain more "difficult" words. (No instances occur in the sample sentence). For example, if one of the ambiguous personal/possessive pronouns is encountered, a check is made to see whether the following sentence element is nominal. If it is, then the pronoun is joined in the element as a modifier, and the pronoun is treated as possessive. Forms of the verb быть which were not covered by the provisions of pass five, are also included in the sixth pass.

Having completed the sixth pass, no further analysis of the Russian sentence is undertaken. The sentence structure

delineated by the analysis passes is not complete, since no attempt is made to set up a clause structure. However, in order to facilitate the task of the synthesis procedure, all the separate group structures (and any remaining separate elements) are arbitrarily connected together in one or more higher groups, presenting the appearance of a unified whole to the synthesis stage.

On the other hand if further analysis passes were applied, particularly with reference to clause delimitation (for example, see Appendix 1 of reference 1), then the sample sentence would appear as:-



Thick lines indicate those connections which our analysis routines have created, and the thin lines indicate those which would have been created by additional routines.

The translation sequence is completed by an English synthesis process. This determines re-orderings, insertions, inflections and selections of English equivalents, and, finally, the format of the printed output, produced by the computer on paper tape and printed on a flexowriter. This process is described in the companion paper, which also includes an account of the descriptive model mentioned above.

References

1. McDANIEL, J., DAY, A.M., PRICE, W.L., SZANSER, A.J., WHELAN, S. and YATES, D.M. "Translation of Russian scientific texts into English by computer — a final report". National Physical Laboratory, Autonomics Division report 35, June 1967.

2. National Academy of Sciences/National Research Council, "Language and machines; computers in translation and linguistics". 1966.

3. DAVIES, D.W. and DAY, A.M. "A technique for consistent splitting of Russian words". Proc. Intl. Conf. on Machine Translation of Languages and Applied Language Analysis, H.M. Stationery Office, 1962, 1, 343-362.

4. YATES, D.M. "A computer model for Russian grammatical description, and a method of English synthesis in machine translation". This Conference.

The work described above was carried out at the National Physical Laboratory.