MACHINE TRANSLATION RESEARCH AT THE NATIONAL PHYSICAL LABORATORY, TEDDINGTON

by A. J. SZANSER, M.Sc., F.I.L.

The author wishes to point out that the sample given below is no longer representative. Since the article was written, earlier this year, further progress has been made in the implementation of syntactic features in the Laboratory's programmes. The project has, in fact, now reached the stage of evaluation by invited outside specialist readers who send articles for translation and subsequently comment on the results.

I. Introduction

THE research on machine translation (MT) has been carried on at the National Physical Laboratory (NPL), Teddington, for about six years. It was started on account of the steadily increasing volume of foreign, notably Russian, technical literature,* which to a considerable extent remained out of reach of scientific and industrial research workers. Neither private activities, nor the published 'cover-to-cover' translations of whole journals could cope with the quantity of the material. The difficulties concerned both speed and cost of the output. The NPL project has been specifically confined to translation from Russian into English, and to the field of electronics and allied subjects.

Before reporting on this activity, a word of warning is necessary. Critics of MT usually point out the virtual impossibility, at least at present, of achieving anything approaching perfect translation. From this (correct) opinion it is easy to jump to the conclusion that, therefore, all MT research is a kind of wild goose chase. However, perfection and practicability are poles apart, and their very distance leaves space for justifiable research. The question as to the future possibility of a literary-standard machine translation remains open, but the aim of the NPL research is much more firmly based, namely to provide, quickly and inexpensively, *usable* translations in a limited technical field for a reader who is expert in the field concerned. This definition narrows the scope of the work considerably, at the same time making it easier and, it is hoped, realizable. The research also brings forth some side profits, both in the field of general linguistics and in respect of particular applications which, however, need not detain us here. As far as the writer is aware, the NPL research group is the only one in this country approaching MT in this way. Although the work has not yet been completed, it is now reaching a stage in which the results are going to be tested on larger samples of texts and readers.

* Meaning both scientific and technological literature.

2. Text preparation and dictionary look-up

The first stage in the process of formalized translation is obtaining the equivalents in the target language of the source-language word units (including the composite ones, such as idiomatic groups) and attaching to them all grammatical information necessary for further stages, in a coded form suitable for processing by a digital computer.

The text must have been given to the computer in a coded form. Fully automatic text reading is the object of another NPL research group, but in the meantime, the text is punched on computer cards.*

The cards bearing the encoded text are fed into the computer.† The text is stored on magnetic tape and further processing is entirely automatic. Since a digital computer operates basically on numbers (normally in the binary notation), each word unit is expressed as a number by a combination of 'ones' and 'noughts', and the same applies to grammatical and other linguistic data. All operations are then, in principle, reduced to dealing with numbers expressed in the above way. At the output, the processed text is re-converted to the standard script by an automatic electric typewriter.

The identification of the original words is achieved by splitting each word into stem and affixes by a special procedure, and then matching the stems against the dictionary entries‡ (allowing for the occurrence of variable stems and homonyms) and the affixes against those possible for the given entry.§ When a match occurs, the complete coded grammatical information, together

* The punching is done with the aid of an electric machine with a Cyrillic keyboard, the cards being fed to it continuously. The punching of an average article from a scientific journal (about 3,000 words) takes a few hours; it does not require a knowledge of Russian. The speed of an electronic reader would, of course, be incomparably higher.

† The computer ACE which is used for this purpose was designed and constructed at the NPL some years ago. It contains mercury delay-line storage with a capacity of 40,000 bits of information, backed by four magnetic drums having a total of one and a half million bits, and six magnetic tape decks.

‡ The stem dictionary on the magnetic tape contains about 18,000 entries (the actual number of different word units is somewhat less, about 15,000, since some of them, especially verbs, have more than one stem) to which about 1,000 more are now being added.

§ This procedure is fully described in references [1] to [3].

with one, or more, English equivalents, is attached to each text word unit and the text so enlarged is called the augmented text. The word units for which equivalents have not been found are left as before, with an added code, which at the output stage will cause them to be transliterated into Latin script. In addition, an affix check is made on these words and it may produce some grammatical information, which is also stored with the unrecognized words and can later be used for the syntactic processing.

This first stage of translation may be divided into (1) the text preparation processes, which consist of numbering all word units, sorting them into alphabetical sequence and splitting off the affixes, and (2) the dictionary look-up, *viz.* matching the stems with the dictionary entries in one parallel run of the text and dictionary tapes[4] and re-sorting them into the original order. Special provisions for stem homography and false affixes are built into the matching procedure. Some other routines, such as recognizing idioms (up to five words in length) or providing for stem analysis in case of non-match (for example in the case of the negative prefix "не") are added. The number of English correspondents has been limited to three at the most, which was helped by the restriction of the field and by the inclusion of a separate idiom list.

The stem-affix dictionary look-up is best suited to a highly inflected source language such as Russian. The alternative, i.e. matching against *all* inflected forms, would be preferable, for example, in respect of English as the source language, and has been, in fact, used for this purpose in the U.S.S.R.[5] Also, as distinct from this 'serial access' look-up, described above, some MT systems use so-called 'random access', matching each text word as it comes.[6] The latter procedure is quicker, especially for small lengths of text, but requires much larger high-speed memory storage. The same applies, to a certain extent, to some ingenious improvements in the serial-access system.[7]

The augmented text is finally re-sorted into the original sequence, using for this purpose the serial numbers allotted initially to the text word units, and is then ready for syntactic processing.

3. Preliminary syntactic procedures

All syntactic programmes operate on the sentence (defined here in a formal way as a string of words between two periods) as a whole.* In the short description that follows, the traditional grammatical terms are used throughout, although their ranges are sometimes extended. The two basic syntactic operations consist each in the integrating of one of the two most common word complexes, *viz.* the nominal group and the predicative group. They do not require the full analysis of the sentence and the group building is done in one pass from the left to the right.

There is no rigid definition of a nominal group (or block), but it has been agreed that such a group should include, in addition to the noun itself:

(a) adjectives (including adjectival pronouns, participles and numerals) preceding and in grammatical agree ment with the noun ('modifiers');
(b) adjectives (including participles) in grammatical agreement and nouns in the genitive case, following the noun; the former separated by a comma, the latter not so separated ('qualifiers'); and
(c) some intervening 'neutral' nouns or groups of words, such as adverbs, parenthetic expressions and prepositional phrases.

Any of (a) met in the basic run of the sentence will start a potential block, which can then be terminated by a noun in grammatical agreement; any of (b) will be added to the just completed block, and any of (c) may be accepted into an opened block.* All, or any, of these may apply in combinations, provided that a number of restrictions and special conditions (which cannot be entered into here in detail) are met.† The purpose of integrating a nominal block is threefold: (i) to solve ambiguities inherent in words by taking intersections of grammatical codes for a group, ‡ (ii) to find the places for insertion of possible English prepositions, equivalent to the Russian cases, and (iii) to reduce the number of the sentence components, which is vital for further syntactic procedures. If, in addition, the nominal block is preceded by a preposition which is known to govern different cases and to produce for each case a different English equivalent, then the nominal blocking is a *sine qua non* for elucidating the meaning. Thus, for example, in the Russian phrase:

"с вышесказанной теории (очевидно…)"

the isolated words may be interpreted as follows: " с "—preposition used with genitive, accusative and instrumental cases (the English equivalent being different in each case); "вышесказанной"—adjective, feminine, in genitive, dative, instrumental or locative cases, singular; "теории"—noun, feminine, in genitive singular or nominative/accusative plural. Taking intersection an unambiguous version is obtained (genitive singular) with the English equivalent:

'from above-mentioned theory (is obvious . . .)'

It should be observed that, if the intersection was not found, each possible case would require a specific preposition to be inserted before both the adjective and the noun in the English output.

An analogous treatment is applied to predicative groups. These groups, hinged upon personal ('finite')

* This is an ad hoc and purely pragmatic definition. The number of definitions that have been proposed is much larger than one may expect (sec, for example [8], pp. 9-28).

* Actually (b) and some cases of (c) are dealt with by a later routine for technical reasons.
† The interested reader can find further particulars in 9.
‡ The term ' intersection ' is used here in the set-theoretical sense, as the sum of properties common to the given group.

verbs and short adjectives,* include modifying particles ("не", "бы") and auxiliary verbs, and may be separated by other words. Their analysis provides the basis for the selection (at the output stage) of the corresponding English forms, which may be quite remote from the literal word-for-word versions. Again special cases are subjects of further routines. An example of a predicative group is:

"(КНИЖКИ) НЕ ПЕЧАТАЮТСЯ (НА ЭТОМ ЯЗЫКЕ)"

Here the analysis reveals the third person plural, present tense, and also the reflective affix "-ся", attached to a verb which is marked in the magnetic-tape dictionary as transitive. The verb is preceded by the negative particle "не". The programmed rule converts such a verb into the passive voice. All the elements, therefore, are ready for the English synthesis at the output, which comes out, after re-ordering according to other rules, as:

'(books) are not printed (in this language)'

The two preliminary syntactic procedures, which are described above, complete what has been so far implemented in the set of programmes ready to operate in conjunction with the magnetic-tape dictionary. A sample of the translation produced at this stage is shown in the table opposite. It has been obtained in a fully automatic way from an actual Russian technical text and no 'manual' simulation or editing has been involved. A number of more refined routines has been elaborated and tested on a simulated augmented text, which is far more flexible for this purpose and consists only of a condensed set of grammatical data. The operation of these routines is, however, fully automatic. They are described in the following section.

4. Main syntactic procedures

Once the nominal and predicative blocks have been determined, main syntactic procedures, operating on the whole sentence, are started. If the sentence is a compound one, the first operation is the recognition of the boundaries of its constituent simple sentences, which are termed "clauses". The clause is defined, for this purpose, as possessing no more than one subject and one predicate, either of which may be compound. It is complete if it contains one of each kind in grammatical agreement. Otherwise, it is either elliptic (either the subject, or the predicate is missing) or mixed (subject and predicate do not agree).

The programme selects suitable subject-predicate pairs and searches for the most likely division points between the clauses. These may be provided by subordinate conjunctions, relative pronouns (and adverbs) or commas in certain positions. An example of a short sentence with one clause embedded in another is:

"опыт, физики подтверждают, удался"

Here the first noun is recognized as the potential subject (nominative/accusative singular), the second noun is ambiguous (genitive singular or nominative/accusative plural)*, the first verb is in agreement with the second noun only, second verb with the first only; commas provide the dividing points; the second noun is therefore selected as nominative (there are no transitive verbs and so the accusative cast does not apply for either noun), so that the final translation is:

'experiment, physicists report, was a success'

The next basic syntactic procedure is the co-ordinate group blocking programme, linking together such strings of grammatically homogeneous words as ' oranges and lemons', or 'came, saw and conquered'. The reader may observe here that, unless this is done, the clause determination is bound to go wrong. This would possibly lead to the conclusion that the co-ordinate blocking should come first. This is, however, not possible as two successive and apparently homogeneous terms may, in fact, belong to different clauses, as, e.g. in ' the current enters the upper circuit and the lower circuit is now opened'. Here, the co-ordinate blocking, if done first, would link ' the upper circuit and the lower circuit' and this would disrupt the clause delimitation. The way to deal with such difficulties will be discussed below. The co-ordinate group blocking must be carried out, however, before the third main syntactic programme, that of verb government.

The latter programme finds out the relationship between verbs and their noun complements, which is known as verb government.† This government is expressed by the use of a particular grammatical case of the complementary noun (or a preposition used with it) and may also be classified into strong and weak government. The strong government has very important syntactic consequences, the weak one is optional and may only indicate preference in solving ambiguities. The government relationship also includes, for programming reasons, the subject-predicate link and the infinitive government, in addition to case government.

The determination of government links serves the following purposes: (i) helping to resolve grammatical ambiguities left over from previous procedures; (ii) preventing the irregular insertion of English prepositions (the most frequent instance being 'OF' before a noun in the genitive case, if it follows another noun; if the former is found to be strongly governed, the insertion is cancelled); (iii) helping to analyse the complete sentence in order to provide the basis for re-arrangement, if necessary, at the English synthesis stage.

An example of the operation of the programme is:

"ученикам награды розданы"

The subject-predicate relationship of the last two words overrules here the weaker 'OF'-relationship between

* All other kinds of predicates (long adjectives, nouns in apposition, etc.) are the subject of a further syntactic programme.

* It is, in fact, a case of homography ("физик", 'physicist' and "физика", 'physics').

† In future development any word (besides the verb) may be considered for case government. The application of such a programme would require an extensive supplementation of the grammatical data in the dictionary.

Russian-English machine translation sample, using the preliminary procedures only

(actual computer output)

ЭТИ ВЗАИМНЫЕ ПОМЕХИ ИМЕЮТ СВОИ ПРИЧИНЫ И ЗАВИСЯТ ОТ

МОЩНОСТИ, КОЛИЧЕСТВА И РАССТОЯНИЯ ПЕРЕДАТЧИКОВ ОТ

ПРИЕМНОЙ АППАРАТУРЫ, РАСПОЛОЖЕНИЯ ИХ АНТЕНН, РАЗНОСТИ

ЧАСТОТ ПЕРЕДАТЧИКОВ ИЛИ ИХ ГАРМОНИК ОТ ЧАСТОТ

ПРИЕМНИКОВ И, НАКОНЕЦ, ОТ ИНТЕНСИВНОСТИ ИЗЛУЧЕНИЯ

ПЕРЕДАЮЩИХ И УСИЛЕНИЯ ПРИЕМНЫХ НАПРАВЛЕННЫХ АНТЕНН В

НЕЖЕЛАЕМЫХ НАПРАВЛЕНИЯХ.

ИСТОЧНИКИ ВЗАИМНЫХ ПОМЕХ ПОДРАЗДЕЛЯЮТ НА ДВЕ ГРУППЫ.

THESE MUTUAL INTERFERENCES HAVE OWN REASONS AND
ALSO

DEPEND ON POWER , QUANTITY(S) AND RANGE(S) OF
OUTPUT ALSO

TRANSMITTERS FROM RECEIVING EQUIPMENT, LOCATION OF
ARRANGEMENT

THEIR AERIALS, DIFFERENCE(S) OF FREQUENCIES OF

TRANSMITTERS OR THEIR HARMONICS FROM FREQUENCIES OF

RECEIVERS AND , AT LAST , FROM INTENSITY OF RADIATION
ALSO FINALLY

TRANSFERRING AND AMPLIFICATION OF RECEIVING DIRECTED
ALSO MAGNIFICATION
GAIN

AERIALS IN NOT WISHED DIRECTIONS .
TRENDS

SOURCES OF MUTUAL INTERFERENCES SUBDIVIDE ONTO TWO
FOR

GROUPS.

the first two. The ambiguous second noun is, therefore, resolved as nominative plural and the translation comes out as:

'prizes are handed out to pupils'

(subject comes first, indirect object takes its place after the predicate).

For the full operation of the programme it is imperative that the clause delimitation and co-ordinate blocking programmes should have been done first. It may, however, be objected again that either of the last two may require at least some information concerning the most vital government links. There are two possible ways out of this difficulty. Either each programme will include some essential parts of the subsequent programmes in a simplified pattern of instructions, or the output of the last one in a set will return to the first for another run (iteration method). The solution actually adopted at the NPL is a combination of both, but the iteration will not be resorted to unless there are some indications of an incongruous sentence.

5. Other syntactic procedures

The choice of syntactic problems to be dealt with and the order in which they are tackled is again approached pragmatically. The actual texts are translated, read and commented upon and, in doing so, fresh syntactic problems are brought to light. In this way two further syntactic procedures have been elaborated, the main criterion for their selection being the frequency of occurrence of the ambiguities in question and the improvement in translation if they are, at least partly, resolved. These are: the adverbial ambiguity and the third-person-pronoun ambiguity resolution programmes.

The ambiguous adverb occurs very frequently in the form adverb/short form adjective neuter, which sometimes is further complicated by additional meaning of: impersonal expressions, conjunctions, as well as comparative degrees of adjectives and adverbs in the short form. Thus "ТОЧНО" may mean 'is accurate' (short form adjective neuter), 'accurately' (adverb) or 'as though' (conjunction); "выше" is 'above' (adverb or preposition), 'higher' (comparative adjective or adverb), and so forth. The line of attack is to pin-point the syntactic function of the ambiguous word. Thus if, considered as a predicate, it is in agreement with the otherwise 'unsaturated valency' of a subject, it is assigned the role of the predicate, as, e.g., in "это вполне точно", 'this is completely accurate'. If, on the other hand, there is no *free* subject in agreement and the ambiguous word is not separated by a comma or commas from the remainder of the sentence,* it is an adverb, as e.g., in "это вычислено точно" 'this is computed accurately'.

The third-person-pronoun ambiguities differ from all the others met so far in that the text information determining their meaning within their respective contexts may be derived from a previous sentence or clause. Indeed, in some cases such information can be obtained only from there, and not from within the same clause, for example:

"Я сделал ошибку, сказал студент.
Но учиельница уже заметила ее."

This applies in particular to the personal/impersonal ambiguity. Russian pronouns "он", "она", may be given the equivalents 'he', 'she' or 'it' respectively, depending on whether they stand for a person or a thing. The same applies to all other grammatical cases of these pronouns. The reference is determined normally by the last appearance of a noun in the gender/number agreement, but the problem is by no means trivial, as is shown by the above example, where "ее" is resolved as the impersonal usage 'it', although the *last* appearance of a noun in agreement was "учиельница" a personal noun.*

Two other dichotomies in the semantic field of the third-person pronoun are caused by its attributive versus predicative role (for example "ее" may be rendered in English by either 'her' or 'hers'), and lastly by the fact that some of these pronouns may be used either adjectivally (possessive pronouns), for example "его" meaning 'his' or 'its' (depending on personality), or else nominally (personal pronouns proper) in which case the same word would mean 'him' or 'it' respectively. This requires syntactic recognition as to whether the pronoun in question is used as a noun complement standing alone or as a modifier in a nominal group.

The threefold ambiguity described above can be resolved syntactically, in a large proportion of occurrences, on the same lines as in other syntactic problems. Complete resolution is not always possible and it has been decided that in such cases two or more admissible meanings should be retained and subsequently shown in the output, thus guiding the reader, who is expected to make the final choice. This principle, incidentally, has been used for all unresolved ambiguities, including polysemantic words. There are, of course, cases where on statistical grounds the odds against the occurrence of a particular form or meaning are so great that a preferential choice may be, and must be, made, but these are not so frequent.

6. Syntactic procedures in preparation

In the description of the completed syntactic programmes references were made to certain special cases to be included in further programmes. Work on some of these is now in progress and two of them are described below.

The elliptic clause programme concerns clauses lacking subject or predicate, or both. The true, so-called 'symmetric' ellipsis is the omission of a syntactic component for stylistic reasons, in order to avoid a repetition.†

* In which case it would be a parenthetic expression.

* The resolution here is not arbitrary or fortuitous, but follows a syntactic rule:'the noun must not be in the subject-object relationship with the pronoun that stands for it'.

† Any constituent of a clause may be omitted in certain circumstances. For a fuller account sec 11.

This kind of ellipsis is not, however, relevant to bilingual MT, in particular Russian-English, since literal translation in this case does not affect understanding.

Moreover, although stylistic ellipsis is common in narrative prose and very common in dialogue, it is rarer in technical and scientific texts.

The second type of ellipsis is an inherent characteristic of the Russian language and has no direct counterpart in English; it is, therefore, important that it should be dealt with in translation. The full (hypothetical) clause has to be restored at the analysis stage, so that the correct English synthesis can be made. The typical example is the lack of a copula when the predicative complement in Russian stands alone: "Петров — дирекор.", "Комната ниская".* The English translation would require here the insertion of the copula 'to be' in the appropriate grammatical form: 'Petrov is director', 'Room is low'. Another example is provided by the impersonal form, of a 'finite' verb in the third person, either singular or plural, when the insertion of the corresponding personal pronoun in English is necessary for understanding, e.g. "Говорят, что..." or "кажется" ('They say, that…', 'it seems').

Another general kind of syntactic analytical procedure on which work is now progressing applies to what can be termed 'individual difficulties'. These are connected with certain very common words which may be used with various syntactic functions, and these functions have to be determined by syntactic analysis (which is, as a rule, sufficient for this purpose) before the English synthesis can be attempted. Here belong such common words as: "и", "а", "что" (with its other forms, more particularly "чем"), "как", "же" and the like, about two dozen in all. It suffices to say that, for example, "и" may have no less than seven distinct meanings which may be expressed in English by (1) 'and', (2) 'also', (3) 'even', (4) 'either', (5) 'indeed ', (6) 'both … and ' (in the disjunctive idiom "и... и..."), and finally (7) no translation. All these may occur in addition to various combinations of "M" with other words, which have to be treated as idioms. Fortunately, the above cases are practically always resolvable by syntactic analysis (for idioms by inclusion in the dictionary, sec Section 2). It, therefore, follows that for a smoother MT each word of this kind should be submitted to a short syntactic routine specifically made for it. The resolution, in most cases, will improve the readability in a considerable measure, since experience has shown that confusion most often arises, not with ' lexically charged' words such as noun, verb, or adjective, but with the function words.†

In addition to the above there is further need to work out solutions to some minor problems that have so far been put aside because of their relatively less frequent occurrence, or because their solution would make a rather small contribution to the better understanding of the translated text. Here belong, for instance, separated idioms and conjunctions, certain indirect forms which would require grammatical transformations and so forth. It is obvious that the question, how far syntactic analysis should be carried, must be answered by a pragmatic approach, i.e. by testing the quality of the translation on a representative sample of readers, and controlled by the law of diminishing returns.

7. Analysis and synthesis

Direct rendering from the source into the target language is both possible and practical only at an elementary stage (say, including the routines covered by Section 3, which is illustrated by the sample). Beyond that, it has been found that the more practical procedure is to separate the analysis and the synthesis stages in the respective languages. Quite frequently the amassed data, determined by analysis, have to be revised, or altered, by some additional processing, iteration or otherwise. As the analysis proceeds, the structure of the sentence emerges and the number of components is reduced gradually (the integrated component retaining, at all times, references to its own parts), so as to obtain in the end a tree structure showing all connexions. The individual components arc termed elements, and they compose together a list structure in which each element contains the grammatical data applying to itself as a whole and the addresses of its constituents, if any. The elements which have no constituents are called terminal. The list structure, complete with all grammatical data and addresses, is stored in the computer and serves as the basis for the synthesis.

If the syntactic role of any element, terminal or not, has not been determined, or two possibilities, either of which is admissible, have been found, this is also coded in an appropriate way. The analysis stage is completed when the whole sentence is represented by one inverted-tree structure, with the top element representing the whole sentence and the terminal elements at the bottom, each representing a single item (a word or an idiom group).

At the synthesis stage this structure is followed again, deciding at each junction the proper order of components in accordance with the rules of English syntax, as well as all necessary insertions and inflexions (with due allowance for irregular forms).

The separation of the analysis and the synthesis does not contradict the essentially bilingual character of the NPL machine translation scheme. The analysis concerns only those features which show either morphological or transformational differences between the two languages. It is not exhaustive and does in no case amount to the introduction of a universal 'interlingua'.

The full analysis and synthesis programmes are now being worked out. (The former will include the analytical

* The long form adjective can often be used predicatively—cf. D. E. Rozcntal "Modern Russian usage", transl. from Russian, Pergamon Press. 1963, p.48.

† For a definition of a function word see Fries, l.c., pp. 87 ff.

procedures described in Sections 4 and 5; this scheme is not yet reflected in the sample). It is hoped that, as a result, much better readability will be obtained.

8. Semantic aspects of MT

In order not to enter the wide controversy among linguists, logicians and philosophers, concerning the equivalence and use of the terms 'meaning' and 'semantic', the writer will confine himself to the so-called 'ordinary' usage. This may be based on the standard monolingual-dictionary entries, and retain, in this way, an entirely pragmatic character, at the cost of some unavoidable circularity.*

The aim of semantic investigation, as defined, is to establish semantic relationships which may help to resolve ambiguities left untouched by syntactic analysis. From the start a distinction has to be made between (a) the use of semantics to resolve syntactic ambiguities, and (b) its use to resolve semantic ambiguities. As regards (a), this possibility has already been explored by marking in the dictionary words possessing properties that can subsequently influence or decide the choice of grammatical forms from the syntactic viewpoint. To this category belongs the 'personality ' mark, whose action was seen in Section 5; here also we may include the group of verbs such as "являться", which govern the instrumental case of a noun. This technique may be extended further, but it is rather limited (cf. also [10]). On the other hand, there are not very many clear-cut relationships of this kind which, in addition, would occur frequently enough to warrant their inclusion both in the dictionary coding and in syntactic analysis.

Concerning the resolution of semantic ambiguities ('b' above), the possibility of its application is much greater, yet the relationships concerned are at the same time much more vague and difficult to tackle in a formalized way. In an elementary form it is applied by the very fact of the limitation of the field of discourse, which has been tacitly adopted by many MT research groups. Such limitation will help, for example, in making the choice between the two meanings of "напряжение" which in electronics will be rendered as 'voltage', since it is not likely to be used there in the sense 'tension' (although this is not excluded). Another example is "лампа", whose English equivalent is 'tube' in radio engineering and electronics, and 'bulb' or 'lamp' in other fields.

The proper use of semantic relationships would refer, however, to establishing links between individual words (or, more strictly, lexical items). By semantic link we shall understand a connexion between two such items, † disregarding their grammatical forms and syntactic functions. The connexions may be of various strengths, which can be expressed as a fraction between 0 and 1. In such a case '1' would represent an obligatory link (which would amount to the two words always occurring together) and '0' a pure coincidence. The relationship can be established either 'manually' by dictionary and text study, or automatically, using matrix procedures, such as those employed by the NPL Information Retrieval research group.

The established links may help to resolve the semantic ambiguities, that is if the source language word has several meanings, each possessing a different equivalent in the target language. There is no need to distinguish between various meanings if they all are covered by one target word (or group of words) as well. This restriction obviously does not apply to monolingual information retrieval, and even less so to theoretical research in pure linguistics.*

An example of a purely semantic ambiguity which can be, in theory, resolved by this method is the word "разряд". Here, the field separation will, in principle, disconnect some of the meanings as, for example, in electronics 'discharge', or in mathematics 'rank, division'. If, however, the text is mixed or common, or does not belong to either of those two fields, the separation would not be adequate. In such a case, the appearance of one or more of the established correlative words within the context could help to make the choice. One may, furthermore, combine semantic and syntactic analysis and so reduce the search for correlatives. Thus, considering only modifiers of the word in question, it may be found that qualitative ones are associated with the physical meaning, e.g. "тихий разряд" ('silent discharge'), whereas quantitative ones, more especially ordinal numerals, go together with the other group of meanings, e.g. "первый (второй...) разряд" is equivalent to 'first (second . . .) division'. Even more specifically, if the latter expression is itself a qualifier, it means the rank, as "ученый первого разряда" = scientist of (the) first rank'. This method can be regarded as an extension of idiom identification, without, however, the simplicity of the latter.

It is evident that with the memories and speed of the available computers, semantic analysis cannot be, at present, either general or complete.† No semantic procedures, apart from a few lexico-semantic rules mentioned earlier, have been used in the NPL project; there is, however, no theoretical obstacle to their being introduced.

9. Conclusion

In the opening section the feasibility of MT research was discussed. It was argued that as long as there is a fair chance of providing usable translations cheaply and quickly enough to cope with the influx of the material, this research is both reasonable and practical. The value of an MT system can only be assessed on the right material and with the right type of reader. Before this is done, any 'assess-

* 'Meaning: what is meant'; 'to mean: (of words): signify, impart'; 'semantic: relating to meaning in language' (Concise Oxford dictionary, 5th edn).

f Co-occurring in some arbitrarily chosen text unit. This would correspond to the term 'collocation', as used, e.g., by Firth.

* In this last respect the work carried out in the Cambridge Language Research Unit is of considerable interest; see, e.g. 12 and 13.

† The matrix techniques, mentioned above, arc applied in a limited field and to small samples (e.g. abstracts) only.

ment' on theoretical grounds, by a non-specialist reader using a random sample, cannot be accepted as valid. Acceptability of output from the NPL system was tested at the preliminary stage (covered by Sections 2 and 3 and illustrated by the sample) and the translations were in general found useful. There is a good hope that at the next stage (as described in Sections 4 to 7) both readability and usefulness will be much improved. Moreover, even if the result of a proper test is negative, further research could still be justified if there exist means and approaches not applied before.

With the present rate of growth of technical literature in all languages and disproportionately smaller increase in the number of qualified translators, the point of saturation, reached some years ago, will be left such a long way behind that it would amount to an actual break in communication.* It is not surprising, therefore, that one may often find experts attempting to provide, as it were, for themselves. They do not read the foreign texts in the ordinary sense, but they are conversant enough with the use of dictionaries and grammars to come to some understanding of the text, albeit very laboriously and slowly. In addition, they are handicapped by the very many exceptions and irregularities in the grammatical rules, as well as by the existence of idiomatic expressions. Now, MT at its humblest can do all that, and a lot more (exceptions and idioms being taken care of), but thousands of times quicker. This, being the very lowest and least-assuming interpretation of MT, is still significantly valuable to those who have tried translating for themselves (there are more such people among leading scientists than it is generally assumed).

At the end of this report it would be pertinent to consider some implications of the development of MT with regard to human translation. Let us begin by dispelling the mistaken notion that MT is only a soulless mechanical imitation of the translating process and those engaged in MT research are 'language technicians', as contrasted with *the* linguists, human translators. Machine translation is, however, also a variety of human translation, but geared to cope with a type of text rather than an individual text, and performed with an entirely different set of tools. These differences bring with them acute limitations, but the latter are expected to be outweighed by the quantity and speed of the output.

Criticism of MT does not always arise from lack of faith in its success. An attack from quite a different viewpoint came from no less an authority than Professor Gabor. According to him, MT may well be possible and may be realized in such sufficiently high quality as to affect the work and status of human translators and, therefore, for social reasons, it should not be encouraged.† In the opinion of the writer this is not, however, a real danger. Let us start from the conclusion of the previous argument. The quality of MT is not likely ever to equal that of human translation; but it may still be usable in the restricted sense. MT, considered in this light, will never displace human translation, it will simply complement it.* It will stand in the same relationship to human translation as many mechanized crafts stand to their corresponding arts, namely they satisfy different needs and serve a different set of customers, but both are necessary. It is not only a distinction between the literary and the technical translation, but also within the latter category there may be two kinds of demand and, correspondingly, two levels of translation. For general acquaintance and survey of a given field, MT will provide a large amount of translation, then for the better digestion of selected items human translators will supply more accurate and polished versions. For a competent linguist there will always be an opportunity, whether his tools are pen and typewriter, or electronic computer and magnetic tape. His brain and skill will always be necessary.

The work described in this paper has been carried out by the National Physical Laboratory.

* According to Professor Gabor, the increase in scientific output is even foster than the population explosion! (see [14], p. 196).

† In the lecture, delivered on 10th February 1965 at the NPL.

* Sec also [15], especially the last paragraph.

References

1. Davies, D. W., "The organisation of a Russian-English stem dictionary on magnetic tape", *Language and Speech,* 3. 1960.
2. Davies, D. W. and Day, A. M., "A technique for consistent splitting of Russian words", *"Proceedings of* 1961 *International Conference on Machine Translation of Languages and Applied Language Analysis*", HMSO, 1962.
3. McDaniel, J., *et al., "* Machine translation at the National Physical Laboratory, Teddington, England ", in "*Progress in Machine Translation*", ed. Prof. A. D. Booth, North Holland & Co. (in the press). (This work is already somewhat out of date.)
4. Szanser, A. J., letter to the Editor, *The Incorporated Linguist,* **3** (4), Oct., 1964, p. 118.
5. Hirschberg, Yu. V., "Opyt realizatsii anglo-russkogo algoritma mashinnogo perevoda . . .", *Nauchno-tekhnicheskaya Informatsiya,* **12,** 1965.
6. Bowers, D. M., and Fisk. M. B., "The World's Fair machine trans lator ", *Computer Design,* Apr., 1965.
7. Lamb, S. M., and Jacobsen, W. H., "A high-speed large capacity dictionary system ", *Machine Translation, 6,* Nov., 1961.
8. Fries, C. C., "*The structure of English",* Longmans, Green & Co. Ltd. 1957.
9. *Annual Reports of the National Physical Laboratory,* Teddington, for the years 1962, 1963, 1964, HMSO. (The 1965 Report has since appeared.)
10. Hirschberg, L., "L'utilisation de 1'information sémantiquc dans la choix des unités lexicales" and also Dubois, J., " Résolution des polysémies dans les textes écrits", both papers read at the "*Colloque International de Linguistique Appliquée",* Nancy, 1964. The writer is not aware of the proceedings having yet been published.
11. Leonteva, N. N., "Analiz i sintez russkikh elliptichnykh predlozhenii, *Nauchno-tekhnicheskaya Informatsiya,***11**, 1965.
12. Sparck Jones, K., "Experiments in semantic classification" *Machine Translation,* **8** (3/4), 1965.
13. Needham, R. M., "Applications of the theory of clumps". *Machine Translation,* **8** (3/4), 1965.
14. Gabor, D., "*Inventing the future ",* Secker & Warburg, 1963.
15. Yngve, V. H., "Implications of mechanical translation research", *Proceedings of the American Philosophical Society,* **108** (4), Aug., 1964.