MECHANICAL TRANSLATION

DEVOTED TO THE TRANSLATION OF LANGUAGES WITH THE AID OF MACHINES

VOLUME FIVE, NUMBER TWO NOVEMBER, NINETEEN FIFTY EIGHT

News

The following report was received from Prof. Ichiro Honda of the Department of Psychology at Kyoto University in Kyoto, Japan. The translation has been authorized by staff members of the Institute of which Prof. Honda is a member.

Since finishing electronic computers ETL-Mark III and IV last fall, we have been working on the English-to-Japanese electronic translator and reader for English, and we have just completed pilot models of both machines. With these machines, sentences typewritten in English are automatically translated into Japanese.

The machine is instructed to translate English into Japanese in the following manner.

1. For each English sentence, one Japanese sentence is processed. (Sentence-for-sentence translation.)
2. After each word of a sentence has been identified in the dictionary, a decision is made about the part of speech of the word.
3. Suffixes such as -s denoting plural and -ed denoting past are handled according to suffix-processing orders.
4. We do not work out a corresponding Japanese word directly from an English word, considering required rearrangement of word order.
5. Because there are four kinds of sentences, declarative, interrogative, exclamatory, and imperative and because there are many types of declarative sentences, sentence types and sentence patterns must be identified. This process is called "classification of sentence patterns."
6. To identify the sentence pattern, the subject, verb, object and complements must be sorted out. The logical process is carried on until the fundamental structure of the sentence conforms to any one of the following five fundamental sentence patterns, setting aside modifiers consisting of unneeded words and phrases such as the adjective preceding the noun and the prepositional phrase following the noun.

I. S-ga V-suru.
Kare-ga Benkyo-suru.
(He studies.)

II. S-ga V-da C.
Kare-ga da Gakusei.
(He is a student.)

III. S-ga Vt-suru O-o.
Kare-ga Ai-suru Musume-o.
(He loves a girl.)

IV. S-ga Vt-suru IO-ni DO-o.
Kare-ga Hasso-suru Ototo-ni Kozutsumi-o.
(He sends a parcel to the brother.)

V. S-ga Vt-suru O-o C-ni.
Karera-ga Senkyo-suru Kare-o Shicho-ni.
(They vote him major.)

7. When it happens that there is more than one part of speech for a word, the first part of

speech is tested. If the pattern of the sentence cannot be identified, we proceed to the second. There is only one corresponding Japanese word for each part of the English word. For example, kono for 'this' (adj.) and kore for 'this' (pro.).
8. After the pattern has been identified, ga is added to the subject, -o to the object, and -ni to the complement; the verb is moved to the end of the sentence; and a roughly corresponding Japanese sentence is constructed.
9. The "near sentence" mentioned above is supplemented with pre-excluded modifiers and all words of the original English sentence are inserted. In prepositional phrases the noun is placed before the preposition.
10. Each English word is replaced by the corresponding inflected Japanese word.
11. The Japanese sentence is written entirely in 'Katakana' and a space is left between words as in the English sentence, except the 'Kakujoshis' of -ga, -ni, and -o.
12. A word not stored in the memory is processed as a noun and typewritten just as it was in English.
13. Because 'Kanji' is not used, 'On' phonetic reading is limited to a minimum in the Japanese sentence produced, and 'Yamato-Koto-ba' (traditional Japanese not derived from Chinese) is preferred, for example, saiwai for kofuku and sare for kanojo.
14. For the time being, sentences containing a relative pronoun are not handled in our program,

Outline of the Apparatus.

1. Input: It reads 73 letters, including the capitals, lower case letters, comma, question mark, etc.. A letter consists of 8 bits (units).
2. Output: 75 letters of 'kana' (dakuten, handakuten, and soukon), and the capitals of the Roman alphabet, etc. 8 bits per letter.
3. Letters inside the apparatus: After being referred to the "dictionary" all letters are represented by word numbers, thereby avoiding inconvenient processing of English words of varying lengths such as very short ones (I, is) and much longer ones (dictionary, representative, etc.). This representation is used until just before the Japanese translation.
4. Memory: We installed a. magnetic drum containing 820, 000 bits of information on English grammar, parts of speech, idioms, translation of each word, Japanese grammar, etc. . The information is arranged in 200 columns. Besides these 200 columns, there are 10 extra columns for the various translation operations, and each word can be processed at the speed of 1ms.
5. Outline of the main apparatus (translator): It looks very much like the electronic computer and consists of about 650 packaged circuits patterned after the ETL-Mark IV electronic computer. There are 46 orders for the operations inside the machine.

Soviet Developments in Machine Translation: Russian Sentence Analysis †

T. M. Nikolaeva, Institute of Precise Mechanics and Computing Technique, Moscow, U.S.S.R.

The principles of Russian sentence analysis in MT are discussed. The description of various methods of receiving the grammar and vocabulary information for every word analyzed is given. The syntactic analysis of the Russian sentence is described.

Primary Analysis of the Russian Sentence

THE ANALYTIC PART of MT in translation from Russian is a system of routines that work out the grammatical and syntactic features of the words needed for translation into another language.

The entire analysis of Russian breaks down into two large, independent parts: a dictionary and a grammatical analysis. Each word in the sentence to be analyzed is examined in the dictionary, after which the word with the appropriate lexical information passes on to the grammatical analysis.

This information indicates the part of speech to which the word belongs, the characteristics of that part of speech, and finally the specific morphological properties of the word. For example, inherent features of the noun are: gender, membership in one of three declensions, relationship to the category of animateness, as well as stem type. Characteristics of the verb are conjugation, quality of stem, and the possibility of being transitive.

We have divided words having specific derivational characteristics into separate groups. An indication of the number of a group will constitute the dictionary information about the derivational properties of the word to be analyzed.

Our dictionary is peculiar in that each word includes only information about its grammatical nature and place in the grammatical system of the language as a whole. This is a special type of dictionary, one differing substantially from the familiar kinds.[1]

Our dictionary lacks information about the semantic side of the word, that is, about its particular meaning. This is explained by the unusual role played in our practical work by Russian, which serves as an intermediary language.

Since the analytic part of translation from Russian in our work is the same for translation into any language whatsoever, it would be useless to give the "translation" of a word in the Russian dictionary, inasmuch as it forms the content of the synthetic part of the dictionaries for the various languages.

In order to obtain the desired dictionary information, it is necessary to reduce the word being analyzed to the form in which it may be found in the dictionary, i.e., to the so-called dictionary form.

The dictionary form of the noun is the nominative case, singular number; of the adjective the nominative case, singular number, masculine gender; of the verb — the infinitive; of the numeral and the pronoun — the nominative case. The remaining, invariable parts of speech have only one form, which is also the dictionary form.

---

† The paper has been recommended for publication by the Conference of Young Research Workers of the Institute of Precise Mechanics and Computing Technique, 1957.

1. Cf. L. V. Shcherba, "An attempt at a general theory of lexicography, " Izvestija AN SSSR, Division of Literature and Language, no. 3, 1944.

If a sentence contains a word in the dictionary form, the word receives the sign FS (Forma Slovarnaja 'dictionary form') and for subsequent analysis passes on to the following routines in order to avoid premature conclusions about the "contextual" features of the given word.[2]

For example, the FS of the word dom 'house' may be either nominative or accusative case, singular number, while the FS of the word soldat 'soldier' may be either nominative case singular number or genitive case, plural number. Similar homonymic forms are distinguished in the appropriate routines.

In the routine designed to reduce a word to the dictionary form, the words are handled in accordance with their endings, which in Russian are adequately distinctive for the various parts of speech. The inflectional structure of Russian and the highly developed system of derivational suffixes along with virtually non-existent infixation and little homonymy of inflection contribute to the rather prompt recognition of the part of speech to which the word being analyzed belongs and help to supply it with the form needed to locate it in the dictionary.

For example, if a word has the ending -ymi or -emu in a given passage, presumably it can only be an adjective or participle.[3] Therefore, if after changing the inflection to -yĭ in the first case or to -iĭ in the second case the word still does not appear in the dictionary, the added ending is rejected and the ending of the remaining part checked for one of the participial suffixes. The word is then given the form of a verb, since the infinitive form of the corresponding verb is the dictionary form of a participle. It is less complicated to recreate the dictionary form of a verb found in context in the dictionary form since the endings of verbs in the personal form are almost non-homonymic.

However, inflectional homonymy is a rather complex phenomenon. Even in Russian where it is comparatively slight, it causes substantial difficulties. For example, the ending -i may indicate: 1) the plural number of soft-stem short-form adjectives, e.g., sini 'blue'; 2) the imperative form of verbs, e.g., zhivi 'live'; 3) various cases of the noun koni 'horses', knigi 'book,' etc. In such cases the word undergoing analysis is treated in several blocks successively where the various endings of the dictionary forms are generated that are possible for a given ending of the parts of speech — until the word is found.

Stem alternation in many Russian words constitutes another difficulty in reducing a word to its dictionary form. This is characteristic of verbs e.g., beru-brat' 'to take,' greb-gresti 'to rake,' etc. adjectives (dolog-dolgi 'long,' uzka - uzok 'narrow,') and numerals (vosem' - vos'mi ' eight').

However, despite their seeming variety the number of such variations is rather small, and the alternation affects a limited number of vowels (o-zero, e-zero, '-e, etc.). This makes it possible to use uniform methods of reducing such words to their stem form. In the case of alternation of consonants (archaic forms of the past tense of verbs) the number of such variations is even less since the infinitive of all verbs with irregular past tense endings can end only in -eret' (umer -umeret' 'to die'), -nut' (sokh - sokhnut' 'to dry'), or most commonly, in -sti. Therefore, in the following verb types. greb 'he raked,' nes 'he carried,' or mel 'he swept' ending in sti, the same simple command serves to reduce all these verbs to the dictionary form: "Reject the last letter and add the ending -sti."

Suppletion of individual forms of several verbs, pronouns, and non-pronominal substantives is taken into account by entering the suppletive form directly into the dictionary.

Accordingly, the entire routine breaks down into a number of separate parts more or less corresponding to the division in the parts of speech. We must mention the fact that practical necessity compelled us to make subgroups of nouns with endings in -ie, -iĭ, and -ija in the dictionary form, nouns in -mja, and degrees of comparison in -zhe and -she.

Since homonymic endings are analyzed in several parts, a word passes from one part to another until the final stage. The first attempt at a routine based on the endings themselves proved too clumsy to be practicable.

A word that cannot be found in the dictionary after going through the entire routine remains in Russian letters in the translation.

---

2. "Contextual" features are speech, not language, phenomena — for example, the particular case and number of a noun or tense, voice, mood, number, and person of a verb in every sentence.

3. Whether or not the ending refers to a pronoun (to nemu) is readily detected by checking for initial n before -emu.

The routine examined above makes it possible not only to obtain the dictionary form of a word by which it can be located in the dictionary, but also to find out its various contextual and grammatical features. For example, treatment of the word delaesh' ' you do' permits the word to be located in the dictionary and produces the following tags: second person, singular number,
present tense, and predicate.

Thus, the contextual tags of words with non-homonymic inflections are ascertained in the very first routine.

The contextual features of words with homonymic inflections are ascertained in the subsequent routines on the basis of tactic and syntactic principles of context analysis.

The group of operations used in our work to reduce words to their dictionary form, obtain dictionary information, and analyze non-homonymic inflections is called the "primary analysis."

Determination of Morphological Tags

Certain features of the materials we studied — chiefly mathematical literature — make it necessary to devise a special routine to analyze the function of signs that are not words written in Russian letters. We have agreed to call these signs formulas whether they are formulas in the usual sense of the word or symbols of something in non-Cyrillic letters.

The need to devise this routine arose from the fact that the so-called "formula" is not a "foreign" body within a unilingual flow of speech capable of being mechanically translated from one language into another like chapter numbers, figures, etc., but a full and equal member of the sentence performing the function of some part of speech.

Hence, the purpose of the routine to be described is to determine the part of speech to which any formula encountered in a text may functionally belong. A tactic principle underlies this routine, namely, confirmation of the meaning-differentiating role of word order in ascertaining the part of speech of the invariable word.

If the formula under analysis contains the sign $<$, $>$, $=$, $\neq$, or $\rightarrow$ functioning as a predicate, this formula receives the sign "sentence" and undergoes no further treatment. If the formula is preceded by an adjective or verb, it receives the sign "noun"; by a noun, the sign "invariable adjective." If directly followed by a noun, the formula receives the sign "numeral." After going through this routine, all formulas lacking the sign "sentence" are examined for production of the required contextual signs just like ordinary words.

After being subjected to the two routines mentioned above, all words in a sentence will have an indication of the part of speech to which they belong and dictionary information: words that in the given context have non-homonymic endings will also have some contextual features. Subsequent analysis is to obtain the contextual tags of words with homonymic endings and invariable words and to ascertain the syntactic function of each word.

More than the dictionary information about the words and their ending in context is needed in order to obtain this information. The place of each word in the text as a whole, morphological features of the surrounding words, and syntagmatic connections must also be determined. For purposes of investigation such an analysis requires larger semantically self-contained units than the individual words thus far discussed. We call these units "clauses" (Predlozhenie).

We call "sentences" (Fraza) segments of written text divided by several marks of punctuation. Smaller portions of sentences known in conventional grammar as "subordinate clauses" (Pridatochnoe Predlozhenie) or parts of "compound sentences" (Slozhnosochinenoe Predlozhenie) are, in our terminology, "clauses" proper.

In order to analyze the relations of words within a clause we must first isolate it, i.e., locate the beginning and end of each self-contained semantic unit.

We cannot use the existing punctuation marks as dividers since they do not always indicate the beginning or end of a complete thought. They may accompany so-called "parenthetical" words, which carry no syntactic load, or introduce homogeneous members or even simply stress a given word, as in the case of the so-called "sense dash" (Smyslovoe Tire).

We have therefore divided all the marks of punctuation into two large groups — homonymic and non-homonymic. The non-homonymic marks
are the period, question mark, and exclamation point, which always separate individual sentences or entire segments of meaning. The other marks may separate either whole clauses or individual words.

Accordingly, we have designed for the analytic part of our work a special routine for punctuation mark analysis to isolate clauses as self-

contained units. Each mark is checked for its relation to homogeneous and heterogeneous conjunctions, the presence of parenthesis at the words next to the mark, and the presence of a verb with the sign "personal form" to the right or left of the word in question. Depending on the presence of such signs, each mark of punctuation is provided with one of the following signs: "heterogeneous" (Neodnorodnyĭ) (i.e., introducing a subordinate clause), "parenthetical" (Vvodnyĭ) (introducing parenthetical words, participles, and gerunds), "homogeneous complex" (Odnorodno-Slozhnyĭ) connecting parts of a compound sentence) or "homogeneous simple" (Odnorodno-Prostoĭ) separating homogenous members of a sentence). The break-up of a sentence into its individual units of meaning follows the generation of these signs. Analysis by the routines then continues within the clauses thus obtained.

As mentioned above, several contextual signs are ascertained in the first routine. Verbs, the forms of which are for the most part non-homonymic, get the largest number of tags. Only the voice and mood must be ascertained since the features of tense, number, and person of verbs in the personal form have already been determined by the primary analysis.

Analysis of the mood of verbs presents no special difficulty. The imperative is determined by analyzing verb conjugations, while the endings of the subjunctive mood are identified by the presence of the subjunctive particle -by in the clause.

The most complicated problem is that of formal demarcation of the active and passive voices. We have distinguished two types of passive voice: processual - imperfective aspect (e.g., dom stroitsja 'the house is being built') and resulta—tive - perfective aspect (e.g., dom postroen 'the house has been built'). The passive voice of the perfective aspect, which is formed by the short passive participle of the past tense can be readily distinguished, whereas the passive voice of the imperfective aspect, which is formed by addition of the particle -sja to the personal form of the active voice, is sometimes formally almost indistinguishable from the active voice of verbs that have the particle -sja in the infinitive . Various studies of the problem of voice distinctions in Russian usually refer only to the "polysemy" of the particle -sja, but fail to provide criteria for determining the cases where -sja gives the verb purely passive meaning.

The "classical" passive construction in Indo-European languages is a combination of passive subject with verb in the passive voice and agent subject in the instrumental case (e.g., kniga chitaetsja studentom 'the book is read by the student'). This type of construction is rather uncommon in Russian where the active voice predominates and the absence of a subject is expressed, for example by an indefinite - personal sentence. Moreover, even if such a "classical" passive construction occurs, there are cases where structural and sentence homonymy arise. Let us consider, for example, two structurally identical sentences: mal'chik prichesyvaetsja shetkoĭ 'the boy brushes his hair with a brush" and fraza obrabatyvaetsja skhemoĭ 'the sentence is treated in accordance with the routine.' In the first sentence the predicate has the form of the active voice, while in the second the verb, which is externally similar in form to the other verb, obtains the sign of the passive voice.

We started with the dictionary properties of the verbs themselves in our attempt to solve this problem, dividing all verbs capable of receiving the formant -sja into four main categories .

1) Verbs in which the particle -sja constitutes an integral dictionary feature. These verbs, which have only active meaning, make up a group III-ag (e.g., gordit'sja 'to take pride in,' ochutit'sja 'to find oneself' etc.)[4]

2) Verbs in which the addition of -sja is a method of producing a passive meaning. These verbs make up group III-vg (e.g., vyrabatyvat' - vyrabatyvat'sja 'to manufacture - to be manufactured,' stroit' - stroit'sja 'to build - to be built').

3) Verbs in which the addition of -sja is a method of producing a reflexive meaning. These verbs make up group III-bg (e.g., myt' - myt'-sja 'to wash - to wash ourself').

4) Verbs in which the addition of -sja indicates a total change in lexical meaning. These verbs make up group III-gg (e.g., risovat' - risovat'sja 'to draw - to show off,' tronut' - tronut'sja 'to touch - to spoil').

---

4. The first letter of the symbol for this group represents the serial number, the second, the initial letter of the name of the part of speech. (Transl. note: g = glagol 'verb')

Thus when the verbs of group No. III-ag have the particle -sja they automatically receive the sign "active voice," whereas the verbs of group No. III-vg receive the sign "passive voice."

The main difficulty, therefore, comes from the verbs of group No. III-bg and No. III-gg where the distinction in voice is produced on the basis of an analysis of adverbial complements and adjectives of circumstance and the relationship to the category of animateness in the subject and object of the action.

The various signs for the nouns, as discussed above, are produced in the primary analysis routines. This pertains to non-homonymic inflections. For example, a noun in a clause with the case ending -jakh immediately gets the signs plural number, prepositonal case. A noun with the ending -u gets the appropriate signs as follows: if directly after rejection of this ending the word is immediately found in the dictionary, it is naturally a first declension noun, masculine gender. But in a given context it can represent the genitive, dative, or prepositional (more precisely, locative) case, singular number. If the noun is found in the dictionary after replacing the contextual ending with -o, it is a neuter noun and can immediately obtain the signs dative case, singular number.

If it is necessary to replace the contextual ending with -a in order to find this word in the dictionary, the given word, which is a feminine noun, obtains the signs accusative case, singular number.

However, we are far from being able in all cases to produce the signs by using the primary analysis method. We need a special analysis because of widespread homonymy in the genitive and accusative cases of animate, masculine nouns, homonymy in the nominative and accusative cases of inanimate, masculine nouns, homonymy in the locative and dative cases of certain nouns (e. g., les 'forest' - v lesu 'in the forest,' etc.) It is also difficult to recognize the case of third declension nouns which do not distinguish between the endings of the genitive, dative, and prepositional cases or the case of nouns ending in -ie, -ija, or -ii.

The differentiation of homonymic forms in these nouns is effected by the following analysis. The verbal predicate is checked to see whether it is transitive, for the presence of a certain person and number, whether it belongs to a group governing a specific case, and for its position with respect to the noun being analyzed.

Besides analysis of the predicate a check is made for a preposition from a certain group before the given noun (skipping adjectives and adverbs standing before the noun as well as extended attribute sequences.)

The presence or absence before the word being analyzed of a noun or numeral requiring a certain case is also very important in distinguishing between the accusative and genitive cases.

This series of checks makes it possible in the majority of words to determine quite accurately the case of a noun that has homonymic inflections in the sentence under analysis. To illustrate, we shall describe the handling of the word dom 'house', which in the text is in the accusative case, plural number and preceded by na 'in, to,' (preposition group No. I-vpr).

22/23, 3/ [5] – Check the given noun for rejected ending -a_ or -ja.

23/VII, 24/ – Check the given noun for the sign "feminine gender."

24/XI, 28/ – Check for a noun or numeral before the given word (skipping adjectives and adverbs.)

28/XI, 29/ – Check for a preposition from group No. I-apr before the given word (skipping adjectives and adverbs).

29/30, 27/ – Check the given word for the sign "neuter gender" or whether it belongs to group No. I-as.[6]

30/VI, 31/ – Check for a preposition from group No. I-vpr before the given noun.

VI/0, 0/ – Produce signs "accusative case, plural number" for the given noun.

---

5. For an explanation of the symbols, cf. D. Ju. Panov, Avtomaticheskii Perevod (Automatic Translation), Academy of Sciences USSR Publishing House, Moscow, 1956, and I. S. Mukhin, Opyty Avtomaticheskogo Perevoda na Elektronnoi Vychislitel'noi Mashine BESM (Experiments in automatic translation with the BESM Electronic Computer), Academy of Sciences USSR Publishing House, Moscow, 1956.

6. The word dom belongs to group No. I-as, which includes masculine nouns ending in -a in the nominative case plural.

The occurrence of homonymic endings is much more frequent in participles than in nouns. For example, feminine adjectival endings coincide in the genitive, dative, instrumental, and prepositional cases. The case of the adjective is determined on the basis of the syntagmatic connections of the given word, i.e., by the noun to which it is related, with cognizance taken of the possibility of a so-called extended attribute occurring before the noun.

The oblique cases of personal and parts of indefinite-personal pronouns are included directly in the dictionary with an indication of the number and case as well as the stem form of the given word since the grammatical significance of these pronominal forms is in most cases expressed lexically.

The case of forms of indeclinable adjectives and nouns are determined after concluding the analysis of the morphological signs of the entire sentence because a knowledge of the case and number of the adjacent words may help resolve this question.

Moreover, the case and number of indeclinable nouns are determined by analyzing the form of the predicate, prepositional government of this predicate, and adjectives and participles modifying the given word.

Syntactic Analysis of the Sentence

In order to conclude the analysis of a Russian sentence we need more than the dictionary information about the words and their contextual morphological features. The correct transmission of the total meaning of the sentence requires information about the function of each word in the expression of the complete thought and interrelations between the members of the sentence. It is precisely this information that is furnished by the syntactic analysis of each word.

As the science of linguistics developed during the past century, two viewpoints on the essence of the language material studied by syntax have become clearly discernible. Supporters of the one theory regard the sentence as a self-contained complex of words united not only by close internal interconnections but by the fact that they belong to the entire sentence within which each word has its strictly determined place.

Adherents of the second theory treat the sentence as an aggregate of groups. The so-called phrases or syntagmas, each of which is an independent linguistic entity susceptible of syntactic investigation.

Such prominent scholars as F.F. Fortunatov, A.M. Peshkovskiĭ, M.N. Peterson, and others in Czarist Russia dealt with the problems involved in investigating phrases. Their work was continued by A. A. Reformatskiĭ, O. S. Akhmanova, and other Soviet linguists. In recent times the theory of word combinations has been steadily kept in the foreground of linguistic research.[7]

The clear distinction between phrases and sentences, already set forth in the works of the well-known German linguist J. Ries, was continued and extended by Acad. A. A. Shakhmatov.[8]

The compilers of the Academy's Grammatika Russkogo Yazyka (Grammar of the Russian Language) devoted one volume in the section on "Syntax" to an analysis of the types of phrases [9] and another to an analysis of the kinds of sentences.[10]

Development of the new branch of science known as automatic translation, which demands of linguists working in this field maximum precision in defining linguistic categories and concepts, has raised anew the question of the proper unit of context analysis — phrase or sentence.

It is our view that the sentence should constitute the unit of investigation, even though a study of word interconnections in a phrase is of great interest, but, it seems to us, for other purposes.

The focus of attention in our work is the sentence as a whole, which is analyzed not as an aggregate of individual syntagmatic patterns but as a complex entity with individually interrelated parts.

---

7. Cf. F.Mikush,"A discussion of structuralism and the syntagmatic theory, " Voprosy Jazykoznanija (Problems in Linguistics), 1957, No. 1.

8. J. Ries, "Was ist ein Satz?" Prague, 1894; J. Ries, "Was ist Syntax?" Marburg, 1931.

9. Mention should be made of the fact that the concept of phrases held by the compilers of the Academy's Grammar differs from that supported by representatives of the school of Acad. F.F. Fortunatov.

10. Grammatika Russkogo Jazyka, Vol. II, Syntax, Part I and II, Academy of Sciences USSR Publishing House, Moscow, 1954.

Sentence analysis, therefore, consists of isolating the principal members of the sentence (subject and predicate) and the secondary members . We have included among the secondary members, objects, attributes, and adverbs, although, as A.B. Shapiro correctly points out in his interesting article, this "classical" threefold division does not correspond to the complex interrelations arising between the members of a sentence.[11]

Most textbooks of Russian cite the "free" word order in the Russian sentence as an indisputable fact. However, examples of such "freedom" do not at all correspond to reality. The order in ja kupil knigu 'I bought a book' cannot be changed into knigu ja kupil without damaging the sentence and ignoring the normative function of the word order, which rests on a centuries-old language tradition. In the second example the sentence tends to enlarge owing to the defining adverb: e.g., kupil vchera 'I bought yesterday,' kupil na ulitse Gor'kogo 'I bought on Gorky Street,' kupil s udovol'stviem 'I bought with pleasure.' Moreover, these textbooks usually ignore the stylistic peculiarities of the various branches of literature, which also make possible a more precise differentiation of the members of the sentence. Thus, the sentence my uravnenie reshili 'We solved the equation,' lit. 'We the equation solved' is wholly acceptable word order in other branches of literature, but is virtually impossible in mathematical texts.

Recognition of the fact that word order in Russian is by no means free, that, on the contrary, it functions largely to distinguish meaning, enabled us to devise a routine to effect the formal isolation of the members of a sentence.

In referring to the rigidity of Russian word . order we have in mind more than the ordinal succession of the members (i.e., which member of the sentence comes first, which second, etc.). It is also important for us to know in what order certain members of the sentence and parts of speech may precede each member of the sentence.[12]

11. "In investigation of secondary members in the Russian sentence," Voprosy Jazykoznanija, 1957, No. 2.

12. Similar positions. Cf. Ch. C. Fries, The Structure of English, Harcourt, Brace and Company, New York (1952).

There were two possible ways of setting up this routine:

1) To search for each of the potential members of the sentence throughout the sentence.
2) To analyze each word in turn.

We chose the second way as being less clumsy.

Before analyzing a sentence the "parenthetical" (Vvodnye) parts are automatically skipped. These "parenthetical" constructions are separated on the basis of information obtained from the "punctuation-marks" routine. Gerunds and participles are analyzed independently.

If the sentence is complex, the remaining portion is broken down into individual, simpler elements; if the sentence is simple, it is treated as a whole. The division also proceeds on the basis of data produced by the "punctuation-marks" routine.

Verbs with the sign "personal form" and modal adverbs of Group No. II-bn, which more or less correspond in traditional terminology to the "category of state" (Kategorija Sostojanija) receive the sign "predicate" (skazuemoe) during the primary analysis. Gerunds receive the sign "adverb of manner" (obstoyatelstvo obraza deistvija) also during the primary analysis. Verbs in the infinitive receive one of these signs: "part of compound verbal predicate" (chast' sostavnogo glagol'nogo skazuemogo), "object" (dopolnenie), or "non-agreeing attribute" (nesoglasovannoe opredelenie), the presence of another verb in a personal form and the contextual-morphological environment of the given infinitive serve as criteria for the choice of sign.

During the analysis of adjectives the presence of nouns syntagmatically related to the given adjectives is revealed. The adjective then receives the sign "agreeing attribute" (soglasovannoe opredelenie). If, however, it does not modify any noun, the adjective passes on to another part of the routine for analysis.

An adjective of this type that does not relate to a noun is first checked for the sign "instrumental case." If the answer is in the affirmative, the verb is checked to see if it belongs in the byt' 'to be', schitat'sja 'to be considered,' etc. group of verbs, which require the predicate in the instrumental case. Verbs of this type are found in group No. II-gg. If the number of this group is indicated in the dictionary information for the given verb, the adjective receives the sign "part of compound predicate" (chast' sostavnogo skazuemogo).

Adjectives without the sign "instrumental case" and not attributes are examined later like nouns.

Nouns to be analyzed are divided into two large groups: with preceding preposition and without preceding preposition. (The possibility of an extended attribute occurring before a noun undergoing analysis is also taken into consideration.)

The analysis within each group takes place by cases. If there is a nominative case, the possibility of two heterogeneous nouns in the nominative is considered. (Should there be more than two such nouns, homogeneous members must necessarily be found among them.)

Depending on the pronominality of one of these substantives, the presence of certain marks of punctuation, or the sequence of words with respect to one another, one of these nouns receives the sign "subject" and the other "nominal part of compound predicate" (imennaya chast' sostavnogo skazuemogo).

A noun in the genitive case may be a non-agreeing attribute or an object. To verify the latter a check is made to see whether there is a negation in the sentence or the predicate belongs to group No. II-ag (verbs requiring the genitive without a preposition.

In determining the function of a postsubstantival noun, it is important to check for the verbal nature of the noun in front of it. If the preceding noun is verbal the noun in the genitive case will always be an indirect object.

Most nouns in the accusative case without a preposition receive the sign "direct object," while nouns in the dative case without a preposition receive the sign "indirect object. "

Substantives in the instrumental case without a preposition receive the tag "adverb" or "indirect object" depending on the pronominal nature of the given substantive, whether animate or not, and its position in the sentence. Our routine for the indirect object expressed by the instrumental case distinguishes between the so-called "instrumental agent" (tvoritel'nyi dejatelja) and the "instrumental instrument" (tvoritel'nyi orudija).

This precise definition of functions of the instrumental case is important in translating, for example, from Russian into English where the two cases will be translated differently.

Nouns with preceding preposition may be defined as "adverbs," "indirect object," or "non-agreeing attributes."

Homonymic endings are analyzed in several parts. Therefore, a word passes from one part to another until the final treatment. Varying this by creating a routine based on the endings themselves rather than on the possible parts of speech proved to be considerably more clumsy.

The routine takes cognizance of the peculiarities of mathematical texts, where adverbs are used as non-agreeing attributes. In such cases the adverb is checked for "quotation marks. " If the answer is in the affirmative, the adverb receives the sign "non-agreeing attribute"; if in the negative, it receives the sign "adverb."

Numerals are equated with nouns in the analysis for function.

Thus, after passing through this last grammatical analysis routine all the significant words obtain syntactic tags indicating their role in the sentence as a whole. The accurate determination of the syntactic function of the words is highly important in translation from Russian.

Creation of an analytic part for translation from Russian testifies again to the complete possibility and feasibility of automatic translation. Furthermore, it seems to us that it is premature to speak of the so-called "limits" of automatic translation since as developmental work proceeds newer possibilities emerge for eventual formalization in language description, even for the use of similar methods to solve stylistic problems.

Research on the structural description of language as a whole is opening up broad avenues not only for perfecting machine translation but also for achieving a deeper understanding of language itself. It is also helpful in looking at a given language from another "viewpoint," requiring a detailed and precise solution of the linguistic problems that arise. The usual statement that various linguistic anomalies exist, without explanation of the cause, is clearly inadequate in our efforts to achieve machine translation.

The method of grammatical and lexical analysis used in MT requires clear criteria for the circumstance producing each linguistic phenomenon.

In conclusion, we should like to list briefly the principal problems that we encountered in our work on a structural description of the Russian language:

1) Problems of word formation.

2) The significance of lexical environment in showing grammatical category.

3) The function of word order in delimiting homonymic inflections.

4) The interrelation and mutual influence of word order, sentence members, and the various parts of speech to which they belong.

5) The problem of equivalence (e.g., nonstandard adverbial correspondences and the verbal noun).

6) Problems related to the specific character of language norms.

7) The problem of lexical units and interlingual phraseological correspondences.

8) Information about scientific literature as an independent linguistic style.

Some Problems in the Mechanical Translation of German

Leonard Brandwood, Birkbeck College, London, England*

I. RELATIVE CLAUSES

The problems discussed are those of syntactical ambiguity and multimeaning in translating relative pronouns from German to English. The former, which is of concern for the English word order, arises from the coexistence in German of homomorphous inflections and variable word order, the latter from this combined with gender dissimilarities in the two languages. Some statistics are given of the frequency with which such ambiguities were encountered in scientific texts, and some possible solutions or partial solutions discussed.

OUR CONCERN will be primarily with the problems of word order and multimeaning, and with these not in all their aspects — which would be too vast a subject for a short article — but only in connection with one particular part of sentence structure, the relative clause.

Besides relative adverbs, such as worin. darin, etc., which cause no difficulty, German uses three words to introduce relative clauses — der, welcher and was. Certain grammatical forms of these are common to two cases, with the result that the syntactical function of such forms is ambiguous, the types of ambiguity being three in number.

1. The masculine singular nominative of der and welcher is identical to the feminine singular dative.

2. The masculine singular accusative of welcher is identical to the dative plural.

3. The nominative form is identical to the accusative in the feminine singular of der and welcher, in the neuter singular of der, welcher and was, and in the plural (all genders) of der and welcher.

The first two types are rare in comparison with the third, the first especially so, because it can arise only when the relative pronoun is not preceded by a preposition. If it is preceded by a preposition, it is thereby denoted as the feminine dative, since no preposition is constructed with the nominative case. On the other hand, without a preposition the feminine dative form is very seldom encountered. For instance, in samples of text amounting to about 30, 000 words it never occurred once, while the masculine nominative occurred over 50 times. In the second type of ambiguity the dative form is likewise rare without a preposition, preference being given to the formally distinct denen. The presence of a preposition, on the other hand, does not solve the problem as before, unless the preposition is of the type which can be constructed either with accusative or with dative, but not with both.

However, we need not continue to discuss the solution of these first two types, since it will be contained in that of the main problem, the distinction of nominative from accusative in the feminine and neuter singular and the plural of all genders.

In English the functions of subject and direct object in a relative clause are indicated by the fact that, when the relative pronoun is the subject, the direct object is separated from it by the verb, e.g.,

Animals which eat men.

When the relative pronoun is the direct object, the subject occurs on the same side of the verb, e.g.,

Animals which men eat.

In German this distinction cannot be made by the position of the verb because of the rule that in a subordinate clause the verb must normally come at the end. How, then, are we to determine the function of the relative pronoun in those instances where the form of the relative provides no help?

---

* Now at the University of Manchester, Manchester, England.

1. The first step is to look at what follows immediately after the relative pronoun, leaving particles and the like out of consideration. If what follows is not a substantive, or if it is, but its form excludes the possibility of it being nominative, then the relative pronoun may be taken to be the subject of the clause. This applies to about half of all the instances where the function of the relative pronoun is ambiguous.

2. On the other hand, if the form of the substantive following the relative pronoun can only be nominative, the relative pronoun must be the direct object. This accounts on average for another 10 per cent of the total number of instances.

3. Thirdly, the substantive following the relative might be either nominative or accusative according to its form, but is indicated as one or the other by its congruence or otherwise with the verb (for this of course the relative pronoun must be identifiably the opposite number in each case). This too applies to about 10 per cent of all instances.

The remaining 30 per cent are those where the functionally ambiguous relative pronoun is followed by an equally ambiguous substantive. It is these which pose the real problem. Consider, for example, the following two sentences

1. Wir werden die Eigenschaften solcher Felder zu untersuchen haben sowie die Bahnen, welche die Elektronen in diesen Feldern beschreiben.

(We shall have to investigate the properties of such fields, as well as the paths which the electrons describe in these fields.)

2. Aus diesem Grunde müssen die Gleichungen in einer Form vorliegen, welche die unmittelbare Verwendung dieser verallgemeinerten Koordinaten erlaubt.

(For this reason the equations must be in a form which permits the direct use of these generalized co-ordinates.)

Now it should be remarked that, in the 70 per cent cases so far solved, the relative pronoun turns out to be the subject in 56 per cent of the instances, the direct object only in 14 per cent. We would therefore expect the ratio to be reversed in the remaining 30 per cent so far unsolved; and the expectation is fulfilled, the relative pronoun being the subject in only 5 per cent, the direct object in 25 per cent. In short, if the machine interprets every functionally ambiguous relative pronoun which it has failed to solve as the direct object of its clause, and adopts the appropriate word order, it will be wrong once in every six such instances. Judging from the frequency of relative pronouns in texts investigated, this would mean about three times in every 10, 000 words. An idea of what the incorrect word order would sound like can be obtained from the verbatim translation of example 2 above.

If a more positive solution is required, it will be necessary to consider not only the relative pronoun, but also the substantive to which it refers . Identification of this substantive alone will sometimes produce a solution by enabling a relative pronoun ambiguous in respect of number as well as case to have its number determined. Provided this differs from that of the succeeding substantive, congruence with the verb will indicate which is the subject, e.g.,

3. Wir werden die Gleichungen in der Form anschreiben, welche sie bei Verwendung dieser Einheiten annehmen.

(We shall write the equations in the form which they assume when these units are used.)

If, on the other hand, the number of the relative pronoun proves to be identical with that of the following substantive, other means of arriving at a solution will be required. A dictionary for the machine must be compiled which classifies words and indicates not only which ones can be constructed together but also in what way. Thus in the sentence

4. Allerdings wird die Wirkung dieser Felder auf Elektronen, welche sie zu verschiedenen Zeiten durchlaufen, verschieden sein.

(To be sure, the effect of these fields on electrons which traverse them at different times will be different.)

"electrons" can "traverse" "fields, * but not vice versa: nor, since Wirkung may be the other word referred to by welche or sie, are "electrons" likely to "traverse" an "action." The possibility, an "action" "traverses" "electrons, " is at once excluded by congruence with the verb, and so on. Similarly in sentence 3 "the equations" may "assume" a "form, " but not the other way round.

How such a classification can be achieved, and, if it is achieved, whether it will provide the complete solution, are questions still to be answered.

Finally there is the question of how the various relative pronouns are to be translated.

Was may also be used to introduce a substantival clause, including direct and indirect questions, in which case its translation is always "what." It would therefore save the trouble of having to make a distinction between this and its use as a relative pronoun, if the latter too could be translated as "what." This is possible, however, only when was refers to a preceding das, though this is the most frequent type, accounting on average for about two-thirds of all instances. If was is translated as "what," the das is left untranslated. Alternatively the das can be translated by "that" and the was by "which," e.g.,

5. Auch in diesem Falle ist es notwendig bei dem anzusetzen, was als das Kernmotiv des Werkes erkannt wurde.

(In this case too it is necessary to begin with what was recognized as the central theme of the work.)

With all other types the relative was must be translated by "which, " and consequently the relative distinguished from the substantival use. This is easy enough if the was is a direct interrogative — not because of the question mark at the end of the sentence, since a relative was might well occur in an otherwise interrogative sentence, but because the direct interrogative use will occur in a main clause, the relative in a subordinate clause. The problem is how to distinguish the relative from the indirect interrogative and non-interrogative use. If the clause introduced by was is the first in the sentence, was is substantival, e.g.,

6. Was Joseph zu tun hat, ist dasselbe.

(What Joseph has to do is the same.)

When the was clause is not the first in the sentence and is substantival, it can usually be recognized as such by the absence in the preceding clause of a neuter substantive to which the was could refer. This is not an infallible rule, however, because the relative may refer not to any particular word but to the preceding clause as a whole e.g.,

7. Der Wettlauf mußte unterbrochen werden, was sehr bedauert wurde.

(The race had to be interrupted, which was greatly regretted.)

A more, but not completely certain solution results from consideration of the fact that being substantival the was clause is therefore a constituent part of an adjacent clause.

Thus, for example, in the sentence

8. Für alle, die diese Ordnung vertreten, ist das entscheidend, was die Existenz dieser Gesellschaft auszeichnet.

(For all who stand for this (social) order what distinguishes the existence of this society is decisive.)

the main clause has a construction which normally requires a predicative, and this is supplied by the was clause.

On the other hand, in the sentence,

9. Unter diesen beiden Bestimmungen läßt sich alles zusammenfassen, was für die alte Generation charakteristisch ist.

(In these two definitions can be comprehended everything which is characteristic of the old generation.)

the construction of the main clause is complete without the was clause, which is thereby denoted as relative.

With welcher too it is necessary to distinguish the relative from the interrogative use, since the interrogative form is always translated by "which," the relative either by "which" or "who". The solution is similar to that for was, however, and need not be repeated.

The main problem with both welcher and der is to decide whether they are to be translated as "who" or "which." This can be done, of course, only by establishing whether the noun referred to denotes a person or a thing. As was mentioned earlier, the relative pronoun can only refer either to the last substantive occurring before it, or — if this substantive is a dependent genitive or part of a dependent prepositional phrase — to the substantive governing this. If there is more than one dependent prepositional phrase, and if these as well as the governing substantive have dependent genitives, there will be several substantives to which the relative

pronoun might refer. Such a collection is not common, however. In most instances — about 90 per cent according to our experience — there is only one substantive for the relative pronoun to refer to, with the result that there is no problem. Nor is there if there is more than one substantive, but all denote either persons or things. The problem arises only when there is a mixture of persons and things, and then only if the substantives concerned are equally capable of being referred to by the relative pronoun, having regard to gender and number.

In this latter case there are two possible solutions. If, as we previously suggested, the dictionary incorporates a system indicating which words are constructed together, reference to this will probably decide which of the substantives, when substituted for the relative pronoun, is appropriate in the context of the relative clause. Alternatively, if we are prepared to accept some loss in variety of expression plus an occasional odd-looking, but not unintelligible translation, the whole problem can be obviated by using the word "that" for all instances of der or welcher, when used as relative pronouns. Or rather all instances except those in the genitive case. When this is possessive, that is to say; when the relative pronoun governs a noun, it can always be translated by "whose," no matter whether referring to a person or thing. When, however, as sometimes happens, the genitive case does not indicate possession, but merely arises from construction with a preposition or verb governing the genitive, it is to be translated in the same way as other instances by "that."

If the relative pronoun is preceded by a preposition, it can still be rendered by "that, " the preposition then being placed immediately after the verb in the English - "This is the man with whom I went" - "This is the man that I went with." It is with those instances, however, that the occasional odd-looking translation mentioned will be likely to arise.

Where the translation "that" fails is in the non-restrictive relative clause, e.g., "Mrs. Smith told Mrs. Jones, who then went and told Mrs. Evans." In German this is common enough with was, but not with der, the similar use of which is frowned upon by some grammarians. Apart from this and one or two other exceptions, such as the case where the substantive referred to is a person's name, the translation "that" is applicable.

II. PREPOSITIONAL PHRASES

The following is a brief consideration of the difficulty in German of determining mechanically whether a prepositional substantival phrase after a substantive is dependent on it or not, the solution of which is essential for correct word order, and therefore in many cases for the meaning in the English translation. As with the relative pronoun, the conclusion to be drawn is that a complete solution to the problem is not possible solely by syntactical considerations.

THE PROBLEM discussed in the preceding section, that of identifying the word to which the relative pronoun refers, leads to the further problem of distinguishing independent and dependent prepositional phrases.

Generally speaking, if the prepositional phrase preceding the relative pronoun is independent of the substantive in front of it, then the relative will refer to the substantive in the prepositional phrase,

1. Sonst müßte die Hochfrequenzkurve oberhalb des Sprungpunktes mit der unteren Kurve übereinstimmen, welche die Abhängigkeit des gemessenen Gleichstromwiderstandes von der Temperatur angibt.

(Otherwise the high frequency curve would have to coincide above the spring point with the lower curve, which shows the dependence of the measured direct current resistance on the temperature.)

If, on the other hand, the prepositional phrase is dependent on the preceding substantive, the relative pronoun may refer either to the substantive in the prepositional phrase, as in

2. Hieraus läßt sich ferner die ursprüngliche Zusammensetzung des Urans und das heutige Verhältnis von Pb/U und Th/U in den Primärgesteinen, die als Muttergestein der Bleiminerale gelten, in guter Übereinstimmung mit den für Granite experimentell gefundenen Zahlen berechnen.

(From this, furthermore, it is possible to calculate the original composition of uranium and the present proportion of PB/U and Th/U in the primary rocks, which are considered to represent the parent rock of the lead minerals, in close agreement with the figures found by experiment for granites.)

or to the substantive preceding the prepositional phrase, as in the following sentence:

3. Da in der Lichtoptik es rotationssymmetrische Anordnungen von brechenden Flächen sind, welche die Abbildungen vermitteln, werden wir unser Augenmerk auf rotationssymmetrische elektrische und magnetische Felder richten müssen.

(Since in optics it is the axially symmetric arrangements of refracting surfaces which mediate the images, we shall have to direct our attention to axially symmetric electric and magnetic fields.)

It might be thought that the relative pronoun here would refer to "Flächen" rather than the abstract "Anordnungen," but this is by no means certain, as may be seen from the following example:

4. Außer den γ-Strahlen ist noch eine neue Art von Teilchen vom Atomgewicht l vorhanden, welche die beobachteten Protonen durch elastischen Stoß auslöst.

(Besides the γ-rays there is present a new kind of particle of atomic weight 1 which releases the observed protons by elastic collision.)

It is not so much in connection with relative clauses, however, that the distinction of independent from dependent prepositional phrase is important, as in connection with word order. In translating, for instance,

5. Wir haben darauf hingewiesen, daß die Laplacesche Gleichung für die elektronenoptischen Felder gegenüber den lichtoptischen Medien eine Einschränkung bedeuten.

the English word order varies according to whether neither, one, or both prepositional phrases are interpreted as dependent on the preceding noun, "equation": — the different versions are

1. We have referred to the fact that the Laplace equation signifies a limitation for electronoptic fields in comparison with optical media.

2. We have referred to the fact that the Laplace equation for electronoptic fields signifies a limitation in comparison with optical media.

3. We have referred to the fact that the Laplace equation for electronoptic fields in comparison with optical media signifies a limitation.

This problem arises in fact only in subordinate clauses and in the part of a main clause after the finite verb. Since in a main clause, unless it is interrogative or imperative, the finite verb must normally be the second syntactical unit, it follows that any prepositional phrase following a substantive which occurs before the verb forms a single unit with this substantive.

The most obvious method of dealing with change of word order, first proposed by Oswald and Fletcher, is to have on the English side of the program a prescribed sequence for the various syntactical units. Basically this is

- (P) S V OP -

(where each of these — excluding the verb — comprises all its dependent units — prepositional phrase, genitives, etc. *) Such a scheme

* The P in parentheses indicates that if a prepositional phrase occurs before the subject in the German, it is to be retained in the same position in the English translation.

will suffice for the majority of clauses to be translated, and, if so desired, the exceptions can be made the subject of subsidiary rules prescribing alternative syntactical patterns. The result may be a somewhat stereotyped word order in the English, but this is no great detriment in translating scientific texts, which in the German itself — as one would expect — tend to have a less varied and less complicated clause structure than in other literature. Hence in most cases the only change necessary is for the subject to be brought before the finite verb in translating a main clause with inverted order, or for the finite verb to be advanced from the end of the clause to a position immediately after the subject in a subordinate clause: that is, the sequences (P) V S O P and (P) S O P V are to be altered to that prescribed.

If this is the limit to the rearrangement of the word order, the problem of the dependent prepositional phrase will apply only to those dependent on the subject, since they are the only ones liable to be separated from their substantive by the verb. It might be thought that it would also apply to those dependent on the direct object, when this occurred at the end of the clause instead of in its more usual place immediately after the subject or the verb — that is in a sequence such as S V P O P. In accordance with the prescribed sequence the direct object has to be transferred to the position immediately after the verb. It is unnecessary, however, to determine whether the prepositional phrase following the direct object is dependent on it or not, because in either case it can be transferred along with it. Sometimes it is not desirable to follow the prescribed sequence in such instances, but this is a separate problem and does not depend on the status of the following prepositional phrase.

There are occasions, however, when it is necessary to determine whether a prepositional phrase after a direct object,is or is riot dependent. These arise with verbs of perceiving and certain others such as "permitting," when the substantive which is the direct object of the main verb is also the subject of the infinitive. An illustration of this is provided by the following sentence:

6. Wir lassen eine beliebige Ebene durch die Symmetrieachse mit der x, z - Ebene des rechtwinkligen Koordinatensystems zusammenfallen.

(We let an arbitrary plane through the axis of symmetry coincide with the xz plane of the rectangular co-ordinate system.)

If we consider only the prepositional phrases which follow immediately upon a substantive, and these only in a subordinate clause or the part of a main clause after the finite verb, then they are more often dependent on the substantive than independent, the proportion being approximately 5:4. In the case of those that are dependent, the substantive on which they depend is — on an average —

> in 20 per cent of the instances the subject,
> in 25 per cent of the instances the direct object,

and in 45 per cent of the instances an independent prepositional phrase.

In the remaining 10 per cent of the instances the substantive is a predicative, an apposition, indirect object, etc. This means that, if the rearrangement of word order is restricted to the subject and to the direct object in the accusative and infinitive construction just mentioned, only about 1 in 9 of all the prepositional phrases following a substantive causes difficulty. On the basis of texts examined this is approximately 6 per 1000 words. If, however, we wish to change the order of the prepositional phrases themselves — if, for instance, in translating sentence 3 we wish to emphasize the last prepositional phrase and say

> "Since in optics it is the axially symmetric arrangements of refracting surfaces which mediate the images, it is to axially symmetric electric and magnetic fields that we must direct our attention."

we shall have to examine not one or two, but all of the prepositional phrases in the sentence concerned.

Even if we adopt the easier course and have to decide whether the prepositional phrase is dependent or not only once in 9 instances, the question still remains of how this is to be done. A partial solution — investigations suggest that about half of the relevant instances may be solved — can be achieved by including in the program various makeshift rules such as the following:

If the subject is followed by a prepositional phrase but the direct object is not, make the construction passive, so that the direct object becomes the subject and the subject the agent, e.g.,

7. Im Falle b) erreicht das Potential auf der Achse im Punkte S einen Extremwert.

(In case b) an extreme value is attained by the potential on the axis at point S.)

Similarly, where necessary, turn personal constructions such as sich lassen into impersonal ones, thereby again making the subject the direct object, e.g.,

8. Wir wollen zeigen, wie sich aus dem einzelnen Lochblendenfeld die Potentialverteilung in einem aus zwei Lochblenden L l und L 2 zusammengesetzten System näherungsweise bestimmen läßt.

(We intend to show how from the single aperture lens field it is possible to determine the potential distribution in a system composed of two aperture lenses L 1 and L 2 by approximation.)

With certain verbs, for instance folgen, when used intransitively or in the passive voice, and providing the syntactical unit preceding the verb is directly dependent on it, the inverted German word order can be retained in the translation, e.g.,

9. Aus den Gleichungen (53) und (55) folgt durch Bildung der Rotation das Gesetz von Biot und Savart in der Form H=.. . .

(From equation (53) and (55) follows by formation of the curl the law of Biot and Swart in the form H=.....)

Likewise when a predicative adjective stands in first position, e.g.,

10. Bemerkenswert ist das Hineingreifen des Feldes durch die Blendenöffnung auf die andere Seite der Blendenelektrode.

(Noteworthy is the intrusion of the field through the diaphragm aperture to the other side of the diaphragm electrode.)

Those prepositional phrases which remain unaccounted for by this collection of rules are best regarded as independent for two reasons, a) because, on an average, of prepositional phrases following the subject only 4 are dependent on it to every 7 independent — that is, of course, excluding those instances where the subject precedes the verb in the main clause, b) because even if the prepositional phrase is dependent on the subject and becomes separated from it in the translation, the resulting word order is in many cases quite normal — as in the translation of sentence 4 (Part I):

> To be sure the effect of these fields will be different on electrons which traverse them at different times.

This scheme, besides providing only a partial solution, lacks uniformity. It would be more satisfactory to have a system of word classification on the lines suggested in the section on the relative pronoun. In this case, however, at least three factors, as well as their relative order, would have to be specified. Thus in

> Allerdings wird die Wirkung dieser Felder auf Elektronen, welche sie zu verschiedenen Zeiten durchlaufen, verschieden sein.

the members of the collocation Wirkung + auf + Elektronen would be denoted as interdependent, whereas in

> Wir werden die Gleichungen in der Form anschreiben, welche sie bei Verwendung dieser Einheiten annehmen.

those of the collocation Gleichung + in + Form would not. Examination of other examples suggests that even this method will not be entirely infallible, but it could be combined with the miscellany of rules previously mentioned and it has the advantage that it is applicable to all prepositional phrases, not merely those after the subject or direct object.

In conclusion it may be said that for relative clauses and prepositional phrases, as for mechanical translation in general, a comparatively few simple rules usually suffice to solve 80 per cent or 90 per cent of any particular problem. The remaining 10 per cent or 20 per cent, however, demands a much greater program for its solution. No doubt it would be possible to work out eventually a complete system — and this we should certainly endeavour to do — but it would be so complex that whether it could be used would depend on how far the design and speed of operation of electronic computers had been or could be improved. Even if a computer with sufficient storage capacity could be built, the price of perfect translation might very well be too high in terms of computer time.

The Use of Statistics in Language Research

A. F. Parker-Rhodes, Cambridge Language Research Unit, Cambridge, England

The literature concerning the application of statistics to linguistic problems and in particular to mechanical translation is reviewed. The conclusion is that much of the work done is of little direct use for mechanical translation, and that some of it is based on a misapprehension of what statistical techniques can in fact do. Statistical methods can play a useful part in the development of mechanical translation procedures once these have been well established, but have little to contribute at the present stage of the work.

THERE ARE many ways in which statistical techniques might be pressed into the service of language research, and in particular the theory of mechanical translation and information retrieval. Most of these have had their advocates, The purpose of this paper is to review briefly the literature of the subject, and to draw conclusions as to how much of this work can be regarded as a legitimate use of statistics, and as to how relevant it is to the progress of language-processing technology.

There appear to be five main topics covered. First, I shall enumerate these, and then I shall refer seriatim to the works available in the C.L.R.U. library upon each of them. 1) Lexicography: this includes the methods and techniques of compiling lexical information, whether this takes the form of a dictionary of a more or less conventional character, or a thesaurus. 2) Approximative Methods: these are methods of machine translation which aim to rely on keeping errors below a preconceived threshold of tolerance; they use statistics mainly to predict how little work need be done to achieve this. 3) Economics: included here are applications of statistics to ascertain the size of computers needed, the time taken to operate programs, etc. 4) Coding: the problems of coding of information have a statistical aspect whenever code-compression is employed. 5) Cryptography: a peripheral subject, but perhaps worth inclusion.

Applications to Lexicography

A good deal of theoretical work has been done on statistical techniques of a kind which could or might be applied to the study of word frequency. The general problems are of a kind of frequent occurrence in biology, and so have received some attention from that quarter. Of this general kind is the work of Good.[1] More specifically concerned with language problems are the contributions of Mandelbrot [2,3] on word-frequencies. This author points out that a knowledge of word-frequency distributions could be useful to the lexicographer, but he is not himself concerned to make this application. In fact, no one seems to have done so, except Koutsoudas,[4] who in fact concludes that the so-called Zipf and Joos laws are insufficient to give reliable predictions of the size of dictionaries needed in machine translation, and consequently recommends the accumulation of further empirical material with this end specifically in view.

---

1. I. J. Good and G.H.Toulmin, "The number of new species and the population coverage, when a sample is increased, " Biometrika, 43, pp. 45-63 (1956).

2. B. Mandelbrot, "Linguistique statistique macroscopique: Theorie mathematique de la loi de Zipf," Institut Henri Poincare, Seminaire de Calcul des Probabilites, (June 13, 1957).

3. B. Mandelbrot, "Structure formelle des textes et communication," Word, 10, pp. 1-27 (1954).

4. A. M.Koutsoudas and R.E. Machol, "Frequency of occurrence of words; a study of Zipf's law with application to mechanical translation, " University of Michigan, Engineering Research Institute, Publication 2144-147-T (1957).

Koutsoudas' statistical techniques are apparently adequate for his purpose, and he has compiled the required data and analyzed them. No one else has apparently taken statistical methods as seriously as this, and most references to the subject merely suggest that an application of statistics to dictionary making should be made,[5] or even in one case that no dictionary could be made without previous statistical analysis.[6]

The use which most of these authors have in mind is to find out how large a dictionary must be in order to contain, with a given fiducial probability, all the words of particular kinds of text. A secondary application is in finding some way of arranging the entries of a dictionary which will reduce searching time by making the most frequent words come up before the less frequent ones. Much more sophisticated is the idea behind compiling a thesaurus. In a thesaurus we have not merely a list of words with coded information upon them, but a mathematical system whose elements represent sets of words, so arranged that, ideally, every word in the system can be defined by listing the sets in which it occurs. If this were done properly, it should be possible to find a word, or at least most words, by specifying not all the sets in which it occurs, but only some of them; thus, it might be possible to specify a set of sets by considering the context of a given word, as well as itself, which would be enough to identify the given word as exactly as we might wish, provided our thesaurus contained enough information suitably organized.

Obviously, the success of such a scheme is a matter which could be statistically assessed, and in some measure no doubt statistically predicted. Thus, those who have considered the use of a thesaurus in MT have not been slow to appeal to statisticians for help in the very considerable labor of compilation involved. However, in fact, they have not progressed very far. As Luhn[7] puts it, "the formation of notational families (his name for thesaurus heads) is a major intellectual effort, to be undertaken by experts familiar with .....the special field of the subject-literature." This major effort has to be done before one can begin to apply one's statistical methods; Luhn himself makes no pretence of actually doing any statistics. On the other hand Gould,[8] who also considers thesaurus methods, presents the appearance of statistical computation. His problem is the translation of Russian mathematical texts into English, and he is concerned to assess the magnitude of the problem of 'multiple meaning' by statistical means. He defines an 'index of multiplicity' in algebraic formulae, and evaluates it for various word-classes (according to the system of Fries[9]), and presents numerical tables of the result. Actually the figures are not statistical in the strict sense, since no significance tests are done (nor is it shown that his index is a sufficient statistic), and the tables only show such facts as, for example, that prepositions are particularly liable to have multiple meanings. It cannot therefore be said that Gould's use of figures has added to what a discursive argument could have more lucidly put across.

One must conclude, from the few attempts which have been made actually to use statistics for lexicographic purposes, that in this field, a valid application exists only after the lexicographic data have been compiled. The same is true, whether the compilation takes the form of a dictionary or a thesaurus. Given these data, one can assess its adequacy, and even propose specific improvements of a major or minor kind, as a result of statistical analysis of its performance. But before the lexicographer has done his work, the statistician has nothing to use as data.

Approximative Methods

One answer to the difficulties raised by the attempt to reduce translation to a mathematically definite procedure is to base one's procedure on the opposite conception, namely that

---

5. N. Chomsky, Syntactic Structures, Mouton and Company, The Hague (1957).

6. V.A.Oswald and S.L.Fletcher, "Proposals for the mechanical resolution of German syntax patterns," Modern Language Forum, vol. 36, no. 3-4.

7. H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information, " IBM Journal of Research and Development, vol.1, no.4, pp. 309-317 (Oct. 1957).

8. R.Gould, "Multiple correspondence," MT, vol. 4, no. 1/2, pp. 14-27 (Nov. 1957).

9. C. C. Fries, The Structure of English, Harcourt, Brace and Company, New York (1952).

that instead of mathematical definiteness one should aim at acceptable approximation to the best that a human translator can do. In that case, it becomes important to know how much work must be directed to removing the errors present in too crude a procedure, in order to reduce the remaining errors to a point below some given threshold of tolerance. This is a statistical problem familiar in industry and in military applications. There seems good reason to expect that, if the approximative approach to MT is accepted as a useful one, it will rest largely on a statistical foundation.

A good example of the kind of work which is relevant to this viewpoint is that of Yngve[10] on 'gap analysis'; even though this is not oriented directly to MT application. This aims to supplement syntactic analysis of a text by a statistical procedure designed to reveal discontinuities between pattern-groups (of words) previously established by analysis of a sufficiently large corpus of texts. Insofar as the results of such analysis can be regarded as an acceptable model of actual linguistic analysis, the procedure is perfectly sound and, it must be admitted, highly ingenious. It is not like the deceptive figuring which we sometimes meet under the guise of statistics in language research. Most often, however, approximative methods are directed to eliminating errors of a lexicographic kind. For example, Glazer[11] has tried to work out the statistics necessary to permit the insertion of English articles into a translation from the Russian. He makes no great claims for the result but it is at least apparent from his work that the amount and detail of the statistical information required to 'solve' this problem, even within the framework of the approximationist philosophy, would be very considerable. In fact, it is unclear why it should be supposed any 'easier' than using real linguistics to do the job.

A better case is made out by King and Wieselman,[12] who have made some useful estimates of the work involved in progressively improving a crude translation by replacing more probable (and thus sooner tried) renderings of a given word or phrase by successively less probable ones. Once again, the conclusion seems to be that an acceptable amount of computation work leads to a still unacceptably erroneous result, though this no doubt depends on the purpose governing our choice of method.

The nature of approximative methods of trans lation is seen at its clearest when the attempt is made to get at the true meaning of a word by comparing it with successively wider areas of 'context.' The idea is that if the word itself is not sufficiently determinate to be translated by one-one equivalence, it may be that comparing it with the next word, or the last word, will suffice to reduce its possible equivalents to one failing that, we try two neighboring words, and so on till the desired result is achieved. This of course is a very crude model of what context really is, and, as I have stated it, depends on the untenable view that each word has a definite number of 'meanings', one of which has to be selected as its translation in the given context. These are just the assumptions made by Kaplan, [13] who made a statistical study of the problem; he collected his data by asking human informants to write down how many 'meanings' of selected words occurred to them, when the said words were presented in company with var ying numbers of neighboring words. His conclusions were not very detailed, largely becaus his informants were too few to provide a really adequate sample, but they showed clearly enough that indeterminacy of meaning was a decreasing function of size of context. There would be scope for a similar study, on a larger scale and with more powerful statistical methods, using a realistic model of what constitutes context and a realistic measure of the indeterminacy of semantic content; this would however be difficult to do. Like most applications of statistics to MT it would only really give useful results when applied to an already mechanized translation procedure. It would be far too slow and laborious to constitute an aid to constructing a mechanized procedure.

---

10. V. H. Yngve, "Gap analysis and syntax," Transactions IRE, vol.IT-2, no. 3, pp. 106-112.

11. S. Glazer, "Article requirements of plural nouns in Russian chemistry texts," Georgetown University, Institute of Languages and Linguistics, Seminar Work Paper MT. 42 (1957).

12. G. W. King and I. L. Wieselmann, "Stochastic methods of mechanical translation," MT, vol. 3, no. 2, pp. 38-39 (Nov. 1956).

13. A. Kaplan, "An experimental study of ambiguity and context," MT, vol. 2, no. *2,* pp. 39-46 (Nov. 1955).

Application to the Economics of Language Processing

It may be objected that it is still much too early to embark on a serious study of the economic aspects of MT. It is necessary, however, from time to time to reassure those concerned that the scale of the enterprise is not wholly disproportionate to the sums which its ultimate users will be prepared to devote to the necessary equipment. It can hardly be said that adequate data yet exist on which to base an informed answer to the question, "How big a computer must one have to do mechanical translation properly?" The question is of course a statistical one and in this sense is relevant to the present enquiry but it need not detain us long. Several workers have referred to the problem, but only Yngve[14] has given any detailed estimates. Their worth is somewhat dependent on accepting a particular view of the nature of the MT procedure but may be accepted to an order of magnitude, at least until more substantial data are available.

Coding and Code Compression

In large measure the coding problems arising in MT and in library work are the same as those occurring in other branches of communication engineering. The need for code compression perhaps arises more urgently in MT, because of the great bulk of the material to be stored, but the mathematical problems it presents are the same as in other fields, except where, as in the use of thesaurus methods, the mathematical structure of the information to be coded imposes special restrictions.

I do not intend to refer to the already considerable literature on code compression. Specific applications to MT have been discussed by Mooers.[15] This work however depends on using a tree-type semantic classification, as has hitherto been done in most information retrieval systems. The statistics of the process would be appreciably different in a lattice system.

Less specific to our immediate subject are the methods, many of them well known, for compressing alphabetic codes. Quite powerful methods are possible here because of the very great redundancy in alphabetic writing. They are discussed, in general terms and without statistical analysis, by Mukhin[16] and Panov.[17] In general it may be said that none of this work is either controversial or novel; but the statistics of code compression in thesaurus systems is still (as far as published work goes) an unexplored field.

Cryptography

As for coding problems, there is a large literature on cryptography and code design which I do not intend to explore. There are however some special points of contact between cryptography and language research in which statistics could play a part. Yngve[18] has written an interesting paper in which he treats of the translation problem (especially translation out of unknown languages) as a special case of the problem of decoding a message without the advantage of a complete code-book to do so. The approach potentially involves the use of statistics, and, while Yngve does not carry the analysis far enough to make actual calculations it is clear that this could be done. The difficulty is that the analogy between translation and the decipherment of a coded message is really more metaphorical than strictly formal. It is therefore unclear how far the results of such investigations will really be relevant.

General Commentary

Of the two main ways in which statistics can be applied to scientific enquiry, the observational and the predictive, only the first has

---

14. V. H. Yngve, "The technical feasibility of translating languages by machine," Transactions AIEE, Paper 56-928 (1956).

15. C. N. Mooers, "Zatocoding and developments in information retrieval," Aslib Proceedings, vol. 8, pp. 3-19 (1956).

16. I. S. Mukhin, An Experiment in Machine Translation Carried out on the BESM, Academy of Sciences of the USSR, Moscow (1956).

17. D. Panov, Concerning the Problem of Machine Translation of Languages. Academy of Sciences of the USSR, Moscow (1956).

18. V. H. Yngve, "The translation of languages by machine," Information Theory, (Third London Symposium), Butterworth's Scientific Publications (London), pp. 195-205.

really been explored in our field. Observational statistics requires that there be a population of entities of which we cannot hope to acquire a complete knowledge, although we can obtain such knowledge of small samples of the population. These samples have to be taken subject to certain rather rigid precautions and in most statistical work are either created by carefully designed experiments or obtained by properly planned observations on the population as it exists in nature.

In the lexicographic applications these prerequisites are not very well met. When the population is the words in a dictionary, it is not a population of which our knowledge is fragmentary in the sense required. On the contrary, we already know (or someone must know) everything about them that we shall ever discover by our analysis, else the dictionary could not have been written. When the population is composed of words in a text, we are in no better position, for although here a real population exists, we either sample the whole population, in which case what we do is not really statistics but census-taking, or we postulate the existence of a population of which our text is a sample. This is in fact what most of the workers along this line appear to do, but it embodies a statistical fallacy, namely, that of creating a sample by definition. It is legitimate to define a population, ostensively or otherwise, and then set about obtaining samples from it, for then the legitimacy of the sampling procedure is open to test and discussion; it is not legitimate to ostend a sample and say "let there be a population of which this is a sample," for then there is no sampling procedure, and the assumptions of probability theory, on which the analysis of the results must be based, will not be correct.

The same objection does not apply to the application of statistics to the study of approximative methods of translation. Here the criticism which suggests itself, against all the work in this field, is the very artificial character of the systems studied. One feels it would hardly be worth while to do very much calculation on such systems. In fact, hardly any has been done. Many have said that they recognize the problem as statistical, but even those who, like Kaplan,[13] actually set out figures do not actually subject them to real statistical analysis. The application of statistics to these approximative methods is still more a potentiality than a fact.

This indeed is largely true of the whole field. There has been far more written about statistical work in translation and information retrieval than actual work done. Apparently no one has yet clearly stated the very limited nature of the applications possible, but many have borne witness to it by inaction. Broadly speaking, the populations which it would be valuable to have information upon are those provided by mechanically translated texts themselves, and the reason that we want to have the information is so as to be able to spot what is wrong with the translation procedure used. Human texts are not suitable material for the statistician because the information we can hope to get from them is either already available or is more efficiently extracted by the methods of the linguist than by those of the statistician.

The indeterminacy which does exist in language is the indeterminacy which arises from the mapping of a continuous territory onto a chart with a finite resolving power; it is not the result of an intrinsically indeterminate use of a discrete set of symbols however complicated. This being so, language can certainly be described in statistical terms. But there is no point in describing it, because the object of the translator (human or mechanical) is instead to use it, in the same sense that one uses a mathematical system to calculate with. Since we shall never do this 'perfectly,' it will always be worth while to estimate the gravity of our failures and this will be a large enough field for the statistician for a long time. But this activity will only begin when the output of failures becomes copious enough to provide the statistician with large populations and the opportunity of applying proper sampling methods to them. This has not yet happened.

Many of those who have written on this subject seem to have the unexpressed belief that there is in language, or our use of it, something essentially indefinite which can be dealt with mathematically only in statistical terms. If this were so, the conveyance of precise information by talking would be impossible. To some extent the area of possible meanings of a remark can be regarded as a probability distribution, but it is of the kind that is almost everywhere zero and has a finite value only within a restricted region. If we deal in 'areas of meaning' instead of in point-like 'right' and 'wrong' meanings, there are indeed definite rules which tell us what remarks do <u>not</u> mean. Deliberately

ambiguous statements can be made in all languages, but even these can be recognized as such by the rules. The problem for the translator is to find out the rules of the languages concerned and to apply them. It is conceivable that this is too difficult for a machine to do; in that case, perhaps a statistical approximation to the desired translation would be a next-best. But it is a substitute, not the real thing.

This paper was written with the support of the National Science Foundation, Washington, D. C.

The following comments were received from people whose work is mentioned in the preceding article. These comments are published with the permission of those concerned.

I agree with the point of view expressed in this paper by Parker-Rhodes, but I fail to see the relevance that he notes of my work on gap analysis to the approximative approach to MT. The gap analysis procedures were intended as a tool for the linguist who wants to discover non-approximative methods in MT.

I would like to see a clear distinction made between analysis of a language for the purpose of deducing its rules or structure, and analysis of a sentence to obtain its structure for possible use when translating it by machine. We may not be able to mechanize the former as easily as the latter. These two kinds of analysis are as different as the science of chemistry, aiming to discover the general laws of chemical composition and reaction, and the analysis of an unknown compound of mixture for its ingredients and their mode of combination.

V. H. Yngve

Footnote 5, and the accompanying sentence in the text (page 2, second paragraph) should be deleted, as factually inaccurate. No such statement is made in Syntactic Structures. Statistics is discussed only on pp. 16,17, — lexicography is not mentioned at all.

Noam Chomsky

I am sorry to say that the wide range of items covered by Parker-Rhodes and the (to me) excessive economy of words made it difficult to follow him in several places, including the section where he deals with my own piece on "Article Requirements of Plural Nouns in Russian Chemistry Texts."

Frankly, I'm not sure that I understand what he is objecting to.

He did not challenge the accuracy or usefulness of the principle of article insertion I proposed or even fault the statistical methodology, as far as I could make out. May I add, for what it may be worth, that I submitted my paper in advance of delivery to a professor of statistics from Stanford, who found my approach wholly acceptable. In the semi-public demonstration of the Lukjanow code-matching technique held in Washington on August 20th, the percentage of correct article placement (in some 300 sentences, including those in the random text) tallied perfectly with the percentage mentioned in my paper. Parker-Rhode's statement "It is unclear why it should be supposed any 'easier' than using real linguistics to do the job" (p. 6) is particularly baffling. Since the article study originated with and was based wholly on an analysis primarily of English usage and possible Russian morphologico-syntactic decision points, and various counts made afterwards only to ascertain whether the formulation provided "useful" predictability, the implication that the tail wagged the dog is certainly unwarranted.

It was not my intention to use statistics to "solve" the problem; rather to indicate that the formulations suggested permit mechanical insertion or omission of articles with a fairly high degree of accuracy. I can't see how statistics as such are useful in MT except as indicators of the validity of a proposed solution.

In my view there is no single solution of a foreign text. Some 15 years experience as a translation editor, translator (both of scientific and

purely literary works), and student of the art of translation have led me to believe that there are likely to be as many versions or solutions of a text (with varying quality, of course) as there are translators. The acceptability of a given translation rests with the individual reader whose reactions are dictated by his background knowledge of the Subject, sensitivity to the nuances of his native language, and the use to which he intends to put the translation. That is why I am a proponent of "approximationism" in language which I think reflects the reality of the human potential, however weak, rather than the ideal, however desirable.

What is needed now as far as the articles are concerned is not more statistical information per se but greater insight into the way they are behaving today. As you know, English article usage has been evolving over a long period of time and the process is far from complete. Under the present influence of the radio and, particularly, the press, with its emphasis on conciseness, there seems to be a trend away from the article in certain types of constructions, e.g. with abstract nouns in possessive phrases. Elsewhere speakers not infrequently have a choice between "a" and "the", etc., with faint semantic or even idiomatic difference between either. How much precision can we (or should we try to) build into a /the translation machine ?

Sidney Glazer

Dr. Gould's untimely and tragic death in the Alps last summer precludes a personal comment on his part. I feel sure, however, that he would wish simply to let his published work speak for itself.

Anthony G. Oettinger

The Storage Problem†

William S. Cooper, Massachusetts Institute of Technology, Cambridge, Massachusetts

The bulkiness of linguistic reference data, contrasted with the limited capacity of existing random-access memory units, has aroused interest in means of conserving storage space. A dictionary, for example, can be considerably compressed, yet at the same time virtually all of its usefulness can be retained. Various approaches to compression are described and evaluated. One of them is singled out for extensive treatment. This approach allows considerable compression of the "argument" part of each dictionary entry, yet it introduces no chance of lookup error, provided the item to be looked up is indeed in the dictionary.

The Storage Problem

A DIGITAL COMPUTER can be used to process a staggering quantity of data. Data that is to be processed needs not tax the memory of the computer, since it can be dealt with a little at a time, and then disposed of. Sometimes, however, the processing itself requires a large store of reference data, and such data must remain accessible throughout the processing — and preferably in the most efficient memory medium available. The mechanical translation process falls into this class; it is inevitable that dictionary or glossary information of some kind must be stored in quantity for reference. Other long tables of linguistic data may also be found useful for translation. The proportion of this reference data that can be stored in the high-speed memory units depends partly on the capacity of the units, and partly on the cleverness of the programmer.

The capacity of most high-speed, random-access memory units which are presently in use for MT experiments is small compared with linguists' needs. Without sophisticated packing techniques, even the information in a small pocket dictionary could hardly be fitted into the high-speed storage of these computers. Special arrangements of the dictionary help (for example, maintenance of a short subdictionary of the most common words in high-speed storage), but it is still necessary to be frugal with memory space. Large capacity, high-speed storage units are being developed, and these should eventually ease the problem, but meantime stop-gap techniques for stretching the effective capacity of existing storage facilities are needed.

The programmer is thus faced with the task of shrinking the dictionary to a minimum volume, without substantially impairing its usefulness. The obvious approach is to attempt to code the data in question into a form that is more compact, but that retains all the original information. An example would be the following rule: "For English, delete every 'u' that follows a 'q'. " Note that this coding process is reversible, for the more compact, coded form may be expanded back to its original form by the rule: "Insert a 'u' after every 'q'."

However, the formulation of rules as simple as the foregoing is highly empirical. Furthermore, simple rules rarely provide a useful degree of contraction. On the other hand, more complex coding operations lead to the ridiculous situation in which storage space equalling that required by the dictionary is needed to encode the material to be looked up or read out. So such recoding approaches, at least at present, seem rather unrewarding.

---

† This work was supported in part by the U. S. Army (Signal Corps), the U. S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S.Navy (Office of Naval Research); and in part by the National Science Foundation.

1. M.M. Astrahan, "The role of large memory in scientific communications," Research and Engineering (Datamation) 4, 34-39 (Nov.-Dec. 1958).

Argument Compression

A more practical approach is to settle for the compression of only part of each entry. The name "argument compression" derives from the viewpoint that a dictionary can be considered as a function. If X symbolizes the word or phrase to be looked up, the dictionary specifies the value of F(X). For example, a French-English dictionary might yield the function value F(X) = "n.,boy" if the argument X = "garçon" were looked up. An entry in the dictionary is thought of as the pair [X, F(X)] for some particular X. Argument compression is confined to whittling down the length of X for every entry.

Although argument compression is a compromise measure, it is nevertheless a very useful one. Certainly in applications where the arguments are long and the function values short, it is most valuable. But even when both X and F(X) are long, argument compression paves the way for some very convenient arrangements. The components of an entry [X, F(X)] may be separated physically in storage, so long as an indication of the location of F(X) is obtained by finding X. (The indication could be the machine address of F(X), which would be stored along with X; or perhaps the location of F(X) could be made derivable from the machine address of X.) In particular, the compressed X's could be kept in core storage, for example, and the uncompressed F(X)'s relegated to tape. In many circumstances, the greater facility with which lookup operations can be performed might recommend this arrangement. Furthermore, a useful element of F(X), such as a part-of-speech tag, might be allowed to accompany X in high-speed storage. If each F(X) comprises several words, it might be practical to list on tape all words appearing in at least one F(X); then F(X) could be indicated by serial numbers accompanying X in core storage. These examples point to the variety of factors that may make argument compression worth while.

Argument compression is unlike the reversible encoding process previously described. All that is required of an argument compression process is that it leave the arguments sufficiently intact to allow one of the entries to be singled out as the correct one. Consequently, a wide variety of devices is available. These devices can be divided into methods that compress each argument individually and methods that compress each argument in a manner dictated by the arguments of neighboring entries.

Suppose that every argument has N characters, or fewer; the first type of device compresses by discarding information from each argument in some ad hoc manner, so that the remainder has the desired length of N' characters. The truncation of every argument after its N^{th} character would be a crude example. Equally unsophisticated would be the removal of some arbitrary portion of each argument, say, every third character. A little better is the system that replaces each argument by its "check sum," which is merely the sum of its characters when the characters are regarded as digits in some number system. In binary computers, arguments must, of course, lie in binary form. One can capitalize on this by forming a "logical check sum"; each argument can be divided into sections of length N', and the logical sum or product of the sections taken. More complicated schemes can be devised at will. In all instances, the X to be looked up must be mutilated in the same fashion as were the entry arguments and then looked up by an ordinary search routine.

In general, automatic dictionaries are susceptible to two kinds of error:

Error 1. When X is indeed in the dictionary, either no value or a mistaken value of F(X) is yielded by the lookup program.

Error 2. When X is not in the dictionary, an F(X) is assigned to it anyway and is, therefore, extraneous.

The compression devices described in the preceding paragraph introduce the possibility of both kinds of error, the reason being that there is no guarantee against two or more different arguments being compressed down to the same form. However, the probability of this happening is surprisingly low[2] if the desired length N' is large enough and if the system of compression is sufficiently "random." If the instances of two arguments being compressed into the same form are few enough, Error 1 can be eliminated by listing the problematic arguments separately in the computer and by checking X against the exceptions list before it is looked up. And there is always the resort of trying slightly modified compression schemes until one that introduces a low error risk is found.

---

2. D.Panov, "Concerning the problem of machine translation of languages, " Publication of The Academy of Sciences of the U.S. S. R., pp. 9-10, 1956.

Such systems have a special advantage: if N' is set equal to or less than the length of a machine address, and every argument can be compressed to length N', then each F(X), or an indication of the location of F(X), can be stored in the register whose address equals the compressed form of X. Not only is the storing of X avoided completely, but the lookup is immediate and involves no trial-and-error system. When data from short dictionaries or subdictionaries is to be stored in a machine featuring multiple address instructions, this arrangement may be ideal.

The second type of device for argument compression depends on some special ordering of the dictionary entries. Then only the relationships between the arguments of succeeding entries need be stored. Here is an instance where the relationships between arguments are so simple that they are known a priori: A table of the cube roots of the positive integers may be stored merely by storing the ascending values of the cube roots in successive registers; the z^{th} register then contains $3\sqrt{z}$, and arguments may be dispensed with.

Unfortunately, dictionary arguments are not as tightly interrelated as numerical arguments usually are. But the imposition of some ordering — say, alphabetic — immediately creates redundancy in the left-hand columns of a list. For example, the following eight words might be found as arguments of consecutive entries in a French-English dictionary:

garçon
garçonnier
garde
gardon
garer
gargantuesque
gargariser
garnir

Only the underlined part of each word differs from its upstairs neighbor. It has been suggested[3] that certain redundant parts of each entry could be deleted and replaced by an indication of the number of letters to be brought down from the preceding entry. For example, this dictionary segment could be stored as:

0garçon
6nier
3de
4on
3er
3gantuesque
5riser
3nir

This representation has the advantage of being reversible, for the dictionary arguments could be reconstructed in full. Neither Error 1 nor Error 2 would occur. The disadvantage of the representation is that the compressed forms are of unequal length, some of them still being very long.

It is a striking and apparently little-known fact that if a word is known to be in the list, it is unnecessary to store anything but the following list, which consists of an indication of the number of letters to be brought down and the first letter of the remainder of each word:

--
6n
3d
4o
3e
3g
5r
3n

Furthermore, if the list is based on the equivalent binary spelling of words rather than on their alphabetic spelling, it is necessary to store only the number of binary digits to be brought down from the preceding entry — the first digit in the remainder is always a one.

The rest of this paper develops the idea and describes the way a word can be looked up in such a list. We call this system "constituent compression." It has the following features:

a) There is no risk of Error 1.

b) It compresses to a high degree. In a binary machine it can shrink an N-bit word down to as few as $N' = \log_2 N$ bits.

c) The lookup method is fairly complicated and slow, although perhaps no more so than the alternative that would be forced by longer arguments . Provision for looking up several words at one time makes the lookup program more efficient.

d) In applications where an Error 2 is possible, the probability of such can be lowered at the cost of retaining, somewhere in the computer, more information from the original argument list.

3. W.N.Locke and A.D.Booth (editors), Machine Translation of Languages, (The Technology Press of M.I.T. and John Wiley and Sons, Inc., New York, May 1955), Chap. 5, "Some problems of the 'word'," by W. E. Bull, C. Africa and D. Teichroew.

Terminology of Constituent Compression

An argument in a dictionary is a string of alphabetic characters, but we must endow it with numerical properties. It is possible to identify each character with a digit in the number system with radix r, where r is at least as large as the number of different characters to be dealt with. But since the argument must certainly become a series of digits when it is placed in storage, it is probably more natural to regard the coded string as the character string. In this case, the radix r would simply be the base of the computer, e.g., r = 2 for binary computers.

Imagine that the arguments are arranged in a vertical list. Append leading zeros to the shorter arguments until all have a common length of N characters. If there are M arguments all told, the list resembles an MxN matrix having the augmented argument A as its typical row:

$$(1)\qquad \begin{aligned} A_1 &= a_{1,1} \ldots a_{1,n} \ldots a_{1,N} \\ A_m &= a_{m,1} \ldots a_{m,n} \ldots a_{m,N} \\ A_M &= a_{M,1} \ldots a_{M,n} \ldots a_{M,N} \end{aligned}$$

The lower-case a's are individual characters which are considered as digits, and a row A is a single number. Our ordering restriction requires that

$$(2)\qquad A_i < A_{i+1} < \ldots < A_j < \ldots < A_{k-1} < A_k$$

under the convention $1 \le i < j < k \le M$.

Next in some number system with radix s (usually s=r), we form a strictly decreasing series of N non-negative integers:

$$(3)\qquad b_1 > b_2 > \ldots > b_n > \ldots > b_{N-1} > b_N$$

When some $a_{m,n}$ from (1) is written after the corresponding b_n from (3), the combination is called a constituent of A_m, and might be denoted $b_n a_{m,n}$ where the conjunction denotes "write end to end" rather than "multiply." When it is not desirable to specify a particular n, C_m denotes any one of the N constituents of A_m. Every constituent can be read as a number in some system with radix as large as the greater of r and s. The expression $C_i \epsilon A_i$ is to be read "C_i is a constituent of A_i." Likewise, $C_i \epsilon A_j$ may be read "C_i is equal to some constituent of A_j." $C_i \notin A_j$ means "C_i is not equal to any constituent of A_j."

For illustration, suppose all characters are decimal digits and r=s=10. Form series (3) from the decreasing integers N through 1. For A_m = 009408, the six consituents of A_m are displayed: $b_1 a_{m,1} = 60$; $b_2 a_{m,2} = 50$; $b_3 a_{m,3} = 49$; $b_4 a_{m,4} = 34$; $b_5 a_{m,5} = 20$; $b_6 a_{m,6} = 18$. These constituents also might have been denoted C_m. The following are true statements: $20 \epsilon A_m$; $18 \epsilon 009408$; $35 \notin A_m$.

A zero constituent is a constituent $C_m = b_n a_{m,n}$ in which $a_{m,n} = 0$. The zero constituents of our example are 60, 50, and 20. The rest are its non-zero constituents.

The distinguishing constituent of A_j with respect to A_i is a function of the two variables, A_j and A_i. But instead of notating it $C(A_j, A_i)$, we write C_j^i, which is intended to suggest that the value of the function is a constituent of A_j. It is defined as the largest constituent of A_j for which no identical constituent may be found in A_i. More precisely, the following two statements hold:

a) $C_j^i \notin A_i$.

b) If $C_j > C_j^i$ for some C_j, then $C_j \epsilon A_i$.

For example, if A_j = 009408 and A_i = 009266, it is found that $C_j^i = 34$, and that $C_i^j = 32$, when (3) is chosen as before.

Under our convention $A_j > A_i$, $a_{m,n} \neq 0$ in any $C_j^i = b_n a_{m,n}$. In the binary number system, $a_{m,n} = 1$ in a distinguishing constituent;

	ARGUMENT MATRIX	CONSTITUENT LIST
A_1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 **[1]** 0 1 1 1 1 0 0 1	8
A_2	0 0 0 0 0 0 0 0 0 0 0 0 0 **[1]** 0 1 1 1 0 1 0 1 0 1	10
A_3	0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 0 1 0 1 **[1]** 0	1
A_4	0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 **[1]** 0 0 0 0 1	5
A_5	0 0 0 0 0 **[1]** 0 1 1 0 1 1 0 0 0 1 0 0 0 1 1 1 1 1	18
A_6	0 0 **[1]** 1 1 0 1 0 1 1 0 1 0 1 1 0 0 0 0 0 1 1 0 1	21
A_7	0 0 1 1 1 **[1]** 0 1 0 1 0 0 0 1 0 0 0 1 0 1 0 1 1 0	18
A_8	0 0 1 1 1 1 0 1 0 1 0 0 0 1 0 0 0 1 0 1 0 1 1 **[1]**	0
A_9	0 0 1 1 1 1 0 1 0 1 0 0 0 1 0 0 **[1]** 1 0 0 1 1 0 0	7
A_{10}	0 0 1 1 1 1 0 1 0 1 0 0 0 1 **[1]** 0 1 1 1 1 0 0 0 0	9
A_{11}	**[1]** 1 0 1 1 1 1 0 0 1 0 0 0 1 0 1 1 1 1 0 1 1 0 1	23
A_{12}	1 1 0 1 1 1 1 **[1]** 1 1 0 0 0 0 0 0 1 0 0 1 1 0 0 1	16

Fig. 1 A binary example of an argument matrix of form (1) with the corresponding constituent list to be stored in its place.

hence it is implicit, and may be omitted. In binary, C_j^i takes the form b_n, which implicitly records the fact that A_j resembles A_i up through its $n\text{-}1^{th}$ bit, and that the n^{th} bit of A_j is different and is a one-bit. We are most interested in distinguishing constituents of the special form C_j^{j-1}; these point out the leftmost digits that differ from their upstairs neighbors in the A_j's of (1). For illustration, these crucial digits are written inside boxes in the binary matrix in Figure 1; a row $A_0 = 0$ is assumed, to provide a definition for C_1^0. The other features of Figure 1 will be touched upon in following sections.

The constituent list, which serves as the substitute for (1) in storage, is a list of M constituents. Its m^{th} member is the distinguishing constituent C_m^{m-1}. This rule defines the entire list except for its first member, C_1^0, which may be taken to be any non-zero constituent of A_1. Figure 1 includes a constituent list. It is written in decimal instead of binary form for readability. To form the list the series (3) was taken to be the integers N-1 through 0. With this choice, the orders of the boxed bits in the argument matrix make up the constituent list. Therefore a statement such as "$C_i^{i-1} \epsilon X$" could be interpreted, "The bit of X of order C_i^{i-1} is a one-bit," in a binary machine.

Lookup Procedure

Assume that X is in the dictionary. This means $X = A_j$ for some value of j. Just what do we seek when we "look up X"? The usual answer would be that the A_m of the argument matrix (1) must be pointed out as an identical match for X. However, under a compression system, the matrix (1) is never actually stored as it stands, so we must be satisfied

with determining the position m of the A_m which would have matched X if (1) had been accessible in its entirety. Thus, in theory, we seek the value of m for which $F[X] = F[A_m]$. For programming purposes, it might be more convenient to handle $F[A_m]$, or its machine address, than to handle m itself.

If (1) could be stored, a slow but sure procedure for looking up X would be to compare X with $A_1, \ldots, A_m, \ldots, A_M$ in turn, and during the process to take note of the position m for which $X = A_m$. But (1) is to be compressed into a constituent list, with the result that direct comparisons are impossible. Therefore, as will be seen later, we have the complication that not just one, but probably many, positions will be noted as the list is swept. Each one is merely a nominee for the desired value of m. After the list has been swept, the choice of the nominees is easily made; the correct one is simply the largest value of m – that is, the value most recently nominated. More precisely, the search must be designed to fulfill two conditions:

Condition I: If $X = A_j$, position j must be nominated.

Condition II: If $X = A_j$, no position m can be nominated if $m > j$.

After such a j is known, the mission is completed, except possibly for gaining access to $F[A_j]$.

The nominator is a block of storage that is set aside for the recording of nominees. If each nominee obliterates the previous nominee while being stored over it, the nominator will automatically display the correct nominee as the search ends.

The carrier is another block of storage that is set aside for bookkeeping purposes. Its contents are excerpts from the constituent list, and change constantly as the search proceeds.

The X to be looked up must, of course, be accessible during the search. If the programmer decides to decompose X into its constituent form, only the non-zero constituents need be available to the search.

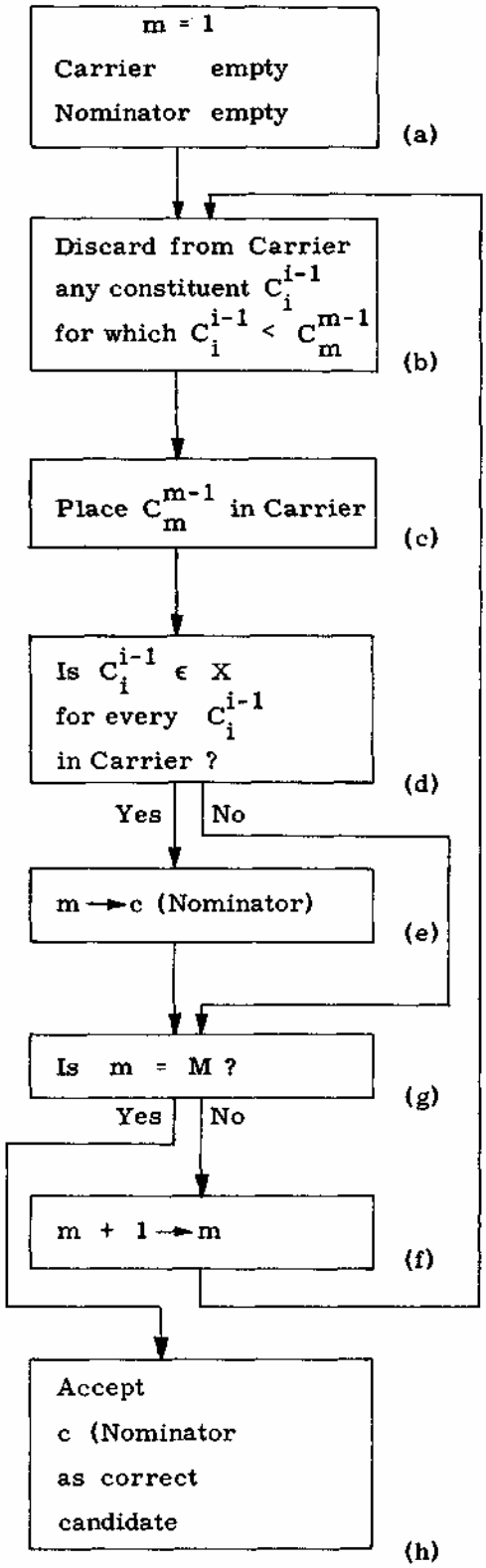

Fig. 2. In this flow diagram, the carrier must be designed to hold as many as N constituents.

The problem: look up X =
0 0 1 1 1 1 0 1 0 1 0 0 0 1 0 0 1 1 0 0 1 1 0 0.

CYCLE	CONTENTS OF CARRIER	NOMINATOR
m=1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0	-
m=2	0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0	2
m=3	0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0	2
m=4	0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0	2
m=5	0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	5
m=6	0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	6
m=7	0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	7
m=8	0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1	7
m=9	0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0	9
m=10	0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0	9
m=11	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	9
m=12	1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	9

The solution: X = A_9

Fig. 3. Contents of carrier and nominator as search of Fig. 2 proceeds down the constituent list of Fig. 1.

There seem to be at least two approaches to performing the search. The first uses a carrier that is equipped to record as many as N constituents at a time. In the second, the carrier contains at most one constituent at a time. The approaches are most easily described and distinguished by means of flow diagrams. They will be discussed in the following two sections.

Search Using a Multiconstituent Carrier

Figure 2 illustrates how a search might proceed. Given the initial conditions of box (a), the loop is traversed M times, one cycle for each successive position m. Boxes (b) and (c) may be regarded as maintenance rules for the carrier, to bring it up to date with m. Box (d) makes the crucial decision of whether or not to nominate the current value of m. An arrow should be interpreted as "replaces, " and c(z) means "contents of z."

A special format for the carrier may be helpful. Let the carrier be simply an N-digit register in the computer:

(4) $d_1 d_2 \ldots d_n \ldots d_{N-1} d_N$

At box (a), every d_n is set equal to zero. In order to place a constituent $C_m^{m-1} = b_n a_{m,n}$ in the carrier, set d_n at the value of $a_{m,n}$ To remove it, set $d_n = 0$ once again. It can be shown that no two constituents need ever share the same d_n in the carrier. The format for the carrier described by (4) allows boxes (b), (c) and possibly (d) to be executed efficiently with shifting operations, especially if the sequence (3) is judiciously chosen so that its members dictate the amount of shift. Also, with format (4), the question of box (d) may be rephrased into a weaker form: "Is each $d_n \leq x$?" where x_n is the n^{th} digit of X.

In a binary machine, format (4) for the carrier may be exploited further. The question of box (d) becomes, "Is $x_n = 1$ for every n for which $d_n = 1$?" Logical operations give a fast answer.

Figure 3 illustrates the problem of looking up X= 001 111 010 100 010 011 001 100 by using only the constituent list in Figure 1. Each line of Figure 3 shows the state of the search after the main cycle of Figure 2 has been performed. The special format (4) has been used to display the contents of the carrier. In place of a value of m, either $F(A_m)$ or its machine address could have been stored in the nominator.

Search Using a Single-Constituent Carrier

If the test of box (d) in Figure 2 remains unwieldy in spite of attempted streamlining, a different approach is needed. Figure 4 displays a search method in which the carrier is never required to carry more than one constituent at a time. Therefore special formats for the carrier need not be devised. Figure 5 illustrates the same problem as did Figure 3. This time, however, the flow diagram of Figure 4 was used for its solution.

Explanation of the Procedures

The lookup procedures of Figure 2 and Figure 4 work on the same principle. Since the binary case is the most easily visualized, we will take as our illustration the argument matrix of Figure 1. Dotted horizontal lines extend from above the boxed one-bits to the right edge of the matrix. Because the list is ordered in ascending magnitude, two little theorems may be proved:

Theorem I: Starting at each boxed one-bit, a "chain" of 1's extends downward until a dotted line is reached (or possibly farther).

Theorem II: Starting just above each boxed one-bit, a chain of zeros extends upward until a dotted line is reached (or possibly farther).

By using the information in the constituent lists, a "cross-sectional" view of the chain of 1's of Theorem I is reconstructed in the carrier for each position m. The search of Figure 2 reconstructs cross-sections of all of these chains (as is apparent in Figure 3), whereas the search of Figure 4 keeps track only of one chain at a time. In either search, every position m is

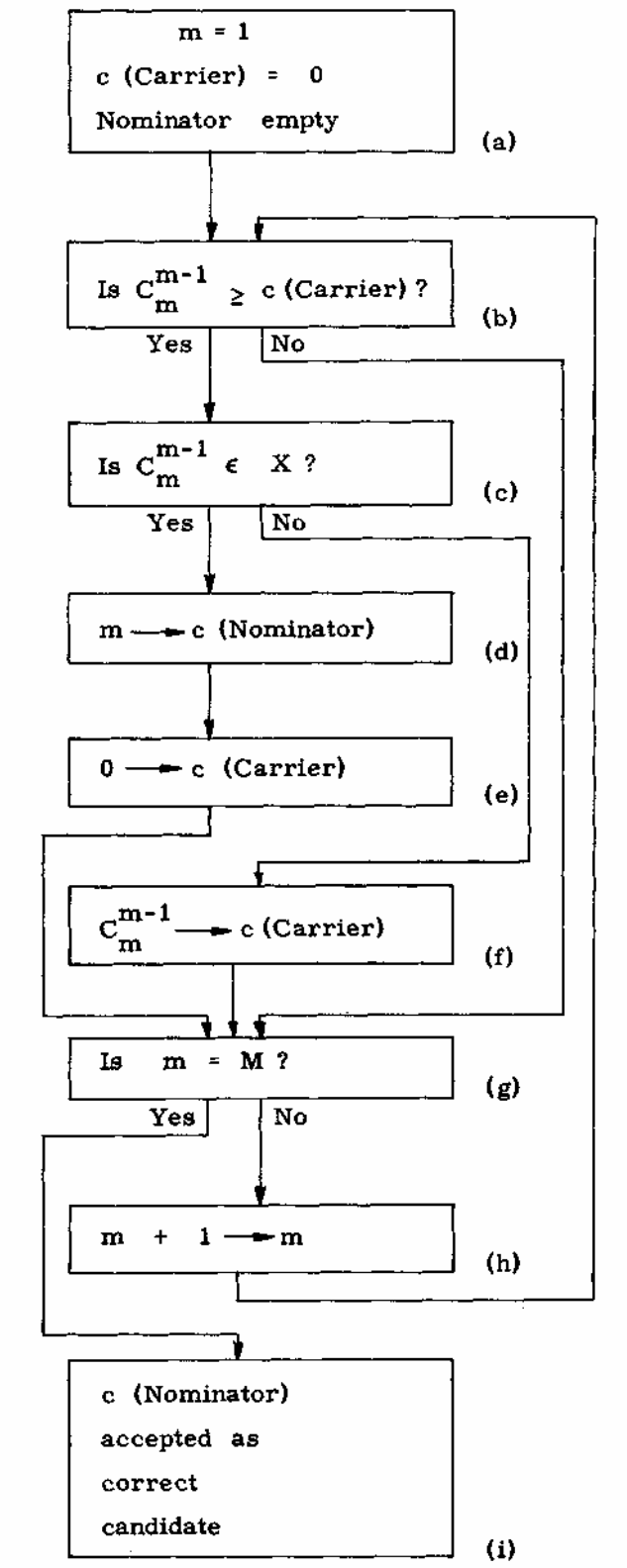

Fig. 4. In this flow diagram, the carrier holds only one constituent at a time.

The problem: look up X =

0 0 1 1 1 1 0 1 0 1 0 0 0 1 0 0 1 1 0 0 1 1 0 0

CYCLE	CARRIER	NOMINATOR
m=1	8	-
m=2	0	2
m=3	1	2
m=4	5	2
m=5	0	5
m=6	0	6
m=7	0	7
m=8	0	7
m=9	0	9
m=10	9	9
m=11	23	9
m=12	23	9

The solution: $X = A_9$

Fig. 5. Contents of carrier and nominator as search of Fig. 4 proceeds down the constituent list of Fig. 1.

automatically nominated, unless the one-bit of some chain in the cross-section for cycle m corresponds to a zero-bit in X.

We know that $X = A_j$ for some j. Condition I, given previously, requires that position $m = j$ be nominated. This will certainly happen, since the 1-chains in the cross-section at $m = j$ are by construction those that are known to pass through A_j. Therefore Condition I is fulfilled, and j is nominated.

Condition II requires that no position $m = k > j$ be nominated. For any position $k > j$, consider the leftmost 1-chain that passes through A_k but that does not extend as far up as A_j. (The search of Figure 2 keeps track of all chains, and hence of this one, whereas the search of Figure 4 is especially designed to keep track only of this chain.) For this particular chain, no dotted lines lie between it and A_j, so Theorem II shows that the one-bit in this chain which lies in A_k corresponds to a zero-bit in A_j above. Position k is not nominated, and Condition II holds. Conditions I and II are necessary and sufficient for the scheme to work.

A rigorous proof is possible. The C_j^i notation is convenient for proofs when $r \neq 2$. The basic stepping stones are theorems that generalize the foregoing theorems. They are stated here in case the reader might enjoy proving them.

Theorem I:

Given matrix (1) under constraint (2). If $C_{i+1}^{i}, C_{i+2}^{i+1}, C_{i+3}^{i+2}, \ldots, C_{j-1}^{j-2}, C_{j}^{j-1} < C_i$ for some C_i, then $C_i \in A_j$ for this C_i.

Theorem II:

Given matrix (1) under constraint (2). If $C_{j+1}^{j}, C_{j+2}^{j+1}, \ldots, C_{k-2}^{k-3}, C_{k-1}^{k-2} < C_k$ for some C_k, and if this $C_k \notin A_{k-1}$, then $C_k \notin A_j$ for this C_k.

Additional Possibilities

So far, it has been assumed that the main cycles of a search must be performed M times. However, a prior knowledge of the approximate position of X down the list is often obtainable; perhaps X is known to match some A_m between A_{m_1} and A_{m_2}. In this case only the relevant range need be searched, so parameters m_1 and m_2 could be inserted in boxes (a) and (g), in either search. If no prior knowledge of the position of X is obtainable, some search time still can be salvaged. Notice that if the largest non-zero constituent of X is exceeded in magnitude by C_m^{m-1}, then neither position m nor any position beyond it is nominable. This statement supplies a stop rule that gives warning when further searching is pointless.

In a scheme that fulfills Condition II, it is plain that if $X_{y_1} > X_{y_2}$, and if position m is a nominee with regard to X_{y_1}, then m cannot be the correct nominee with regard to X_{y_2}. This observation suggests that the X's be ordered according to magnitude prior to the search, the subscripts being assigned so that

$$(5) \qquad X_1 < X_2 < \ldots < X_y < \ldots < X_{Y-1} < X_Y.$$

Then, at any given position m, the X's need be examined in turn only until m is stored as a nominee for some X_y. We now have another stop rule that assures us that the remaining X's may be ignored at position m.

An elaborate but efficient program utilizes both of the preceding stop rules: as m increases, a rising floor value of y is determinable from the first rule, whereas the second rule determines a ceiling value of y at each cycle. Only those X's of (5) carrying subscripts between the floor and ceiling values of y need be considered during any given cycle.

Throughout the discussion, we have assumed that $X = A_j$ for some argument A_j; that is that X is indeed to be found in the dictionary. If we leave the system as it stands, an error of the type described previously as Error 2 is certain to occur whenever a word not contained in the dictionary is looked up. For some special applications, the situation could never arise. With a large enough dictionary, it might arise seldom enough to make the errors forgiveable. Otherwise, it would be necessary to supplement the constituent list with further information about the arguments. A few of the rightmost columns of matrix (1) could be stored, in addition to the constituent list, thereby supplying a few "check digits" for each argument. In order to use the information, the check digits from A_m would be compared against the corresponding digits in X at some stage before $F(A_m)$ could be accepted officially as the correct nominee. The extra information needed might reclaim much of the space saved by compression, but on the other hand, one is free to relegate the check information to a slower storage medium, perhaps along with the F(X)'s. If this sort of error check were programmed, the risk of an occurrence of Error 2 could be reduced to negligible proportions.

I am indebted to V.H.Yngve, K.C.Knowlton, F.C.Helwig, and M. M. Jones for their suggestions and criticism.

Bibliography

153
Linguistic and Engineering Studies in the Automatic Translation of Scientific Russian into English
Technical Report No. RADC-TN-58-321, The University of Washington
ASTIA Document No. AD-148992

This report covers research in machine translation at the University of Washington from May 1956 - June 1958. It includes reprints of publications by members of the group during that period. An outline of the project by the director, Erwin Reifler, defines the problems being considered and the general approach used.

J. R. Applegate

Anthony G. Oettinger 154
An Input Device for the Harvard Automatic Dictionary
Mechanical Translation, Vol.5, no. 1, pp. 2-7

A standard input device has been adapted to permit transcription of either Roman or Cyrillic characters, or a mixture of both, directly onto magnetic tape. The modified unit produces hard copy suitable for proofreading, and records information in a coding system well adapted to processing by a central computer. The coding system and the necessary physical modifications are both described. The design criteria used apply to any automatic information-processing system, although specific details are given with reference to the Univac I. The modified device is performing satisfactorily in the compilation and experimental operation of the Harvard Automatic Dictionary.

Author

H. P. Edmundson and D. G. Hays 155
Research Methodology for Machine Translation
Mechanical Translation, vol. 5, no. 1, pp. 8-15

The general approach used at The RAND Corporation is that of convergence by successive refinements, The philosophy that underlies this approach is empirical. Statistical data are collected from careful translation of actual Russian text, analyzed, and used to improve the program. Text preparation, glossary development, translation, and analysis are described.

Author

Gerard Salton 156
The Use of Punctuation Patterns in Machine Translation
Mechanical Translation, vol. 5, no. 1, pp. 16-24

The determination of sentence structure contributes greatly to the understanding of written texts, and represents, therefore, an element of considerable value in mechanical translation. The present study deals with the analysis of English language punctuation patterns and presents a sample program for an automatic punctuation analysis.

Author

Martin H. Weik 157
Suggestions on a Device for Digital Encoding of Russian Scientific Text
Ballistic Research Laboratories, Memorandum Report No. 1150, June 1958

Various suggestions are made concerning an encoding device for the digital encoding of Russian Cyrillic scientific text. The device is intended as an input unit for mechanical translation equipment.

Author

V. H. Yngve 158
A Programming Language for Mechanical Translation
Mechanical Translation, vol.5, no. 1, pp. 25-41

A notational system for use in writing translation routines and related programs is described The system is specially designed to be convenient for the linguist so that he can do his own programming. Programs in this notation can be converted into computer programs automatically by the computer. This article presents complete instructions for using the notation and includes some illustrative programs.

Author

I. A. Mel'chuk 159
Machine Translation from Hungarian into Russian
Problemi Kibernitiki (Problems in Cybernetics)
No. 1, 1958, Moscow, pp. 222-264

An experimental version of the rules for the machine translation of scientific texts from Hungarian into Russian is described. The article includes extracts from the dictionary of stems, the dictionary of idioms, the list of pecularities, some governing tables (as examples), complete "Tables of Hungarian Suffixes," "Rules for Search, " "Rules for Differentiating Homonyms," and all of the "Analyzing Rules." Several of these are discussed in detail. The importance of a statistical study is suggested.

G. G. Heller

T. N. Moloshnaya 160
Problems in Distinguishing Homonyms in Machine Translation from English into Russian
Problemi Kibernitiki (Problems in Cybernetics)
No. 1, 1958 Moscow pp. 216-221

Lexical and grammatical homonymy is one of the main difficulties in machine translation because it prevents unambiguous determination of the specific class to which a word belongs. Therefore, homonymy must be eliminated before a sentence can be analyzed. This article outlines the system of rules effecting the recognition and elimination of homonymy in order to achieve machine translation from English into Russian. The rules governing the analysis of homonymy have been programmed. Testing on a machine is expected in the near future.

G. G. Heller

Machine Translation (Mashinnyi Perevod) 161
Collection of Articles on Machine Translation
Institut Tochnoi Mekhaniki i Vichislitel'noi Tekhniki AN, SSSR, Moskva, 1958

The papers which appear in this collection were submitted to the Conference on Machine Translation, May 1958, by the linguistic research group of the Institute of Precision Mechanics and Computing Technique of the Academy of Sciences. There is a general theoretical article "Some General Questions in Machine Translation, " by I. K. Bel'skaja. The other articles all of which deal with specific problems are: "Analysis of Punctuation Marks in Machine Translation from Russian," T. M. Nikolaeva; "Translation of Compound Substantives from German to Russian in Machine Translation," C. V. Parshin; "Constitution of a German-Russian Dictionary of Words with Multiple Meanings for Machine Translation, " S. S. Belokrinickaja; "Grammatical Analysis in Machine Translation from Russian to Chinese," V. A. Voronin; and "Some Questions in Machine Translation from Japanese to Russian, " M. B. Efimov.

E. S. Klima

R.H. Richens 162
Tigris and Euphrates — A Comparison between Human and Machine Translation
National Physical Laboratory, Teddington, Middlesex, England, Paper 2-4

After an analysis of the domain of symbolization in terms of categories ranging from initial symbols through mediate symbols to 'terminal indicata,' a comparison is made between the ways in which human and mechanical translation deal with these symbols. In syntactic analysis where MT methods are generally based on principles that have no obvious parallels to human translation, a method based on a multi-stage analysis of lexico-grammatic symbols is suggested. To deal with the semantic aspect of the symbols, their terminal indicator 'naked ideas,' to which human translators must usually resort, a logically formalized standard language as interlingua is proposed. Semantic analysis must follow and be closely linked with syntactic analysis. Three separate operations for semantic analysis are proposed.

J. Viertel

163

Z. M. Volotzkaya, I.N. Shelimova, A. L. Shumilina,
I.A.Mel'chuk, T. N. Moloshnaya
"Concerning a Russian Frequency Dictionary Based on the Material in Mathematics Texts"
Voprosy statistiki rechi, Izdatel'stvo Leningradskovo Universiteta, 1958

Work has been started on a statistical study of a mathematics text of 60,000 words. Each word was listed with governing (preceding) and governed (following) word. Words were classified according to frequency of occurrence with the various parts of speech as "governing" and "governed" words. The frequency of parts of speech and grammatical forms (cases, tenses, etc.) was also determined. Some examples of the findings in numbers and percentages are given.

M. Freeman

I. A. Mel'chuk 164
"Statistics on the Dependence of the Gender of French Nouns on the Endings"
Voprosy statistiki rechi, Izdatel'stvo Leningradskovo, Universiteta, 1958

A statistical study carried out on the Spanish rules for the determination of the gender of nouns is reported. It was found that the rules are 97% to 98% accurate. The analogy between Spanish and French based on their mutual derivation from Latin is discussed. Then three formulations of rules for determining the gender of feminine nouns in French are presented. The best formulation is found to be 94% accurate; the more simplified formulation, about 85% accurate.

M. Freeman

165

Materialy po mashinnomu perevodu, Sbornik I, Izdatel'stvo Leningradskogo Universiteta, 1958

This collection of articles contains reports of work in mechanical translation and related fields. There are several general articles and also articles that deal with specific problems in mechanical translation. Among the general theoretical articles are "The Importance of Mechanical Translation for Linguistics," M. I. Steblim-Kamenskij, in which there is a discussion of the new points of view concerning language which have been fostered by consideration of problems in mechanical translation. In a second article, "Linguistic Problems in the Design of a Machine Language for Information Machines," by V. V. Ivanov, there is a comparison of the problems involved in the construction of information machines with those encountered in machine translation. In a third article by N. D. Andreyev, "A Meta-Language for Machine Translation and its Applications, "the special features of the meta-language which must be used in the several phases of machine translation are discussed.

An article "Concerning the Structure of a Dictionary and the Coding of Information for Machine Translation," by I. L. Bratchikov, S. JA. Fitialov, and G. S. Ceĭtin describes and discusses the special problems of constructing a dictionary for machine translation. There is a series of articles in which the problems encountered in work on various pairs of languages are described. These articles are: "The Root Isolating Program for Machine Translation of Indonesian, " by N. D. Andreyev, B. P. Golovanov, L. I. Ivanov, and A. K. Oglovlin; "Work on a Norwegian-Russian Algorithm for Machine Translation," by V. P. Berkov and M. P. Cherkasova; "Initial Stage of Work on an Arabic-Russian Algorithm for Mechanical Translation," by O. B. Frolova and V. I. Strelkova; "Some Problems in the Design of a Burmese-Russian Algorithm for Mechanical Translation," by N. D. Andreyev, E. A. Zapadova, and O.A. Timofeeva; "Outline of a Program for Morphological Analysis of Russian for Mechanical Translation," by L. N. Zasorina, N. B. Karachan, S. N. Medvedeva, and G. S. Ceĭtin; "Concerning Work on a Hindi-Russian Algorithm for Mechanical Translation," by T. E.Katenina; "Elements of an Independent Analysis in a Vietnamese-Russian Algorithm for Mechanical Translation," by N. D. Andreyev, D. A.Batova, V. S. Panfilov, and V.M.Petrova, and "Concerning Mechanical Translation from Japanese to Russian," by A.A. Babincev and JU. P. Semenishchev.

There is an article "The First Stage of an Indipendent Analysis of the Structure of the Simple Sentence in English," by B. M. Leĭkina, which deals with the determination of parts of speech in a sample text, and the final article "Principles of Design of Electronic Reading Devices," by N. D. Andreyev deals with the special problem of orthographic variants.

E. S. Klima

R. H. Richens 166
Interlingual Machine Translation
The Computer Journal, vol. 1, pp. 144-147

The first part of this paper considers some of the reasons why mechanical translation via a logically formalized interlingua is worth pursuing. The interlingua described consists of a network of bonded semantic elements, the bonds being either homogeneous, corresponding to a generalized notion of qualification or heterogeneous, for dyadic relations. The translation procedure involves a basic program applicable to any input language (P) and any output language
(Q), and P-interlingua and interlingua-Q mechanical dictionaries. The essence of the program is the construction of an array of symbols, grammatical, syntactic and semantic, containing all the information required for translation. The interlingual translation of the input in P is then derived by successive eliminations, usually involving comparisons either across the rows of the array or down the columns. Similar treatment of a second array suffices to translate from the interlingua to the output Q.

Author

167
Experimental Machine Translation of Russian to English
Ramo-Wooldridge, M 20-8U13, Dec. 15, 1958

This progress report describes the research procedures used in developing a computer program to produce a translation of Russian physics texts. The approach described is that of improving by successive modifications a word-for-word translation. The translation rules followed in the program are presented, and an example of machine translated text is included.

J. R. Apple gate

168
M. Masterman, R. M.Needham, K. Sparck Jones
The Analogy between Mechanical Translation and Library Retrieval
Preprints of Papers for the International Conference on Scientific Information, Area 5 pp. 103 - 121

This is a discussion of MT, seen as an assimilation of retrieval procedures based on a thesaurus. After examining work done in the Soviet Union, in order to show that there may be ascertainable principles on which a generalized mechanical translation of grammar and syntax could be based, the thesaurus, considered as the general ordering principle of language, is described, together with the accompanying cross reference dictionary, and the procedures by which these are used in translation are outlined and illustrated.

J. Viertel

V. E. Giuliano 169
An Experimental Study of Automatic Language Translation
Harvard University Computation Laboratory, Report No. NSF-1

This doctoral dissertation deals with three major instruments for research in automatic translation developed at the computation laboratory of Harvard University. The instruments are: a system of manual procedures and computer programs for compiling a Russian-English automatic dictionary on magnetic tape; a computer program that uses the automatic dictionary, and a trial translator or proposed system of computer programs for testing experimental syntactic algorithms. Each of the instruments is described and fully discussed.

J. R. Applegate