[*International Conference on Machine Translation of Languages and Applied Language Analysis*, National Physical Laboratory, Teddington, UK, 5-8 September 1961]

INTRINSIC MACHINE ADDRESSING IN AUTOMATIC TRANSLATION

by
Y. LECERF
(Direction Generale Recherches et Enseignements,
C.E.T.I.S., EURATOM)

I

THE NORMATIVE AUTOMATON AND THE PROBLEM OF HOMONYMS

A) THE PROBLEM OF HOMONYMS

IN order to translate in a correct manner from a given text of one language into another, we are forced to treat and reduce numerous cases of these homonymies of the written language which are called homographies. Amongst the German translations of the French word "facteur", it is not a matter of indifference whether one chooses "Faktor" or "Briefträger". To the form "panse", there corresponds a large number of translations, of which certain ones are far-fetched: verbindet, striegelt, Widersäuermagen, Schmerbauch, etc. A word as simple as "porte" can be the occasion of mis-translations (trägt or Tür). Only the context alone can enable us to choose the correct one amongst translations suggested by the dictionaries.

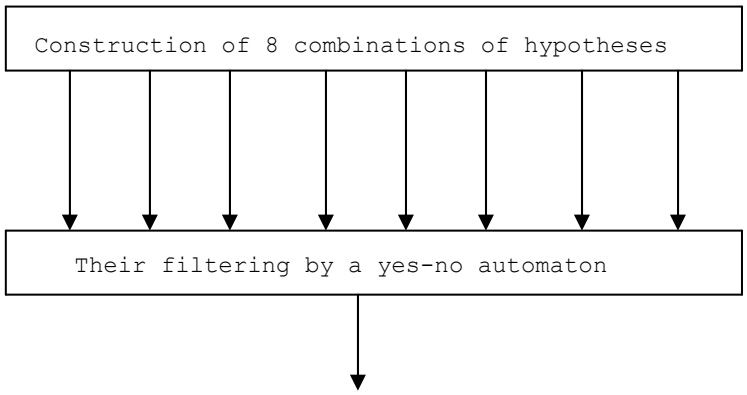
The reasonings which inspired the different methods of homonym reduction are all founded on examining the context of each of the words, whose meanings we wish to decide upon. If the words of the context, words on which we rely for our reasoning, are themselves ambiguous, it is necessary to examine separately each one of the alternatives thus discovered; for example, we may wish to separate out the various following alternatives suggested by the dictionary for the sentence: "les portes-tu?"

les	**portes**	**tu**
Article Pronoun	Verb Noun	Pronoun Past Participle

Although this sentence is very short and while we have mentioned only a portion of the possibilities, the hypotheses given rise to 2 x 2 x 2 = 8 different combinations.

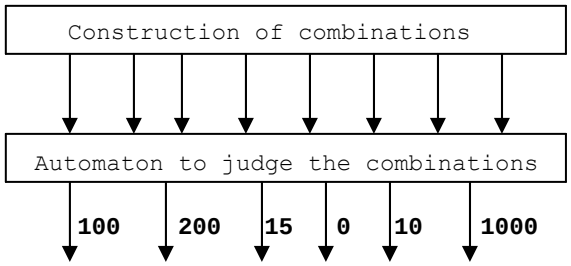
A finite "yes, no" automaton of the Rabin and Scott type (1) produces a general solution to grammatical problems of this kind. If, in receiving a message constituted by a succession of names of grammatical categories, this automaton is capable of refusing sequences which do not correspond to

any correct sentence and accepting the others, then all we have to do is to feed it with these 8 combinations of hypotheses which have been suggested above.



In fact, there exist certain questions to which it is not possible to reply with a definite or distinct "yes" or "no". This is, in part, the case with judgments which may have to be made on certain grammatical structures, a little incorrect perhaps, but which certain authors employ and which everybody understands provided they have no more likely meanings. In the second place, this is the case with all judgments which have *a* semantic background; although the word "crayons" is not an animate being, the one and least likely interpretation which can be put on the phrase, "le crayon éclate de rire", ought to be accepted by the automaton.

As a result of this, there must be a slight modification in its conception. Instead of replying simply by "yes" or "no", this automaton will associate with each of the combinations of hypotheses proposed to it "penalties" or "credits", in proportion as these combinations are more or less likely. The simple "yes" will correspond to a null penalty, a "no" to an infinite penalty. All that remains is to choose, between all the combinations of hypotheses associable with a given sentence, the one with the least penalties.



Thus the possession of a normative automaton, of the kind described above, would be sufficient for resolving the problem of homography, on condition that we had a machine that was sufficiently powerful and that it could work for a sufficiently long time. But is it necessary to have this aptitude for treating the combinations of hypotheses which appertain to the entire spread of the sentence? We can easily show that it is. The homography of sentences such as: "L'obstination de ces homes brave la tourmente", which can be interpreted quite differently from "l'obstination de

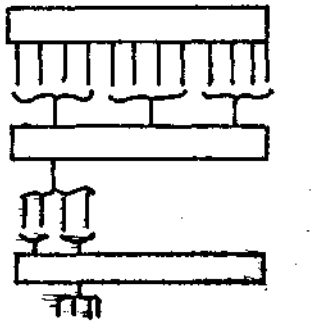
ces hommes braves la tourmente", can only be resolved if we take into account the totality of the sentence. The normal automaton described above thus appears both necessary and sufficient and the theoretical problem of homography is reduced to that of constructing such an automaton.

In practice, one is obliged to focus attention on certain complementary matters which enable us to save machine time. The number of combinations of hypotheses increases, in effect very quickly as a function of the length of the sentences studied. In the example below:-

Le	page	brise	la	pointe	de	la	lance
Art Pr	N.fem. N.Masc.	V N	Art Pr N	V N	Prep	Art Pr N	V N

The number of combinations is 2 x 2 x 2 x 3 x 2 x 1 x 3 x 2 = 288. If we take into account the possibility which exists in French of utilizing each word as a noun, when we wish to designate it as being a word the number of combinations would appear to be very large already. For a sentence of n words, of which each is capable of 1 + k meanings, on the average, this number becomes $(1 + k)^n$, and we see that it would assume enormous proportions in the case of sentences of 100 or 150 words. Thus we make use of a system of prefiltering in groups, in such a manner that the most unreasonable combinations are not even formulated.

Example:



This principle leads us to employ not merely a single automaton, but several automata of increasing selectivity arranged in series. It enables us to take advantage of the possibility of eliminating, at the outset, trials which have to do solely with parts of sentences, for example, detecting impossible binary sequences. All these mechanisms are easy to set up, if we admit that they serve solely to economize in machine time and that the theoretical normative automaton (one which would

judge at a single time each combination) placed in series after them could remedy their eventual insufficiency and guarantee the rigour of the whole operation.

B) TECHNICAL HOMOGRAPHIES

Since, in any case, we have to go to the trouble of setting up a mechanism which is sufficiently good for treating rapidly and with precision the problems of homonymy, it is natural to try to use this mechanism also for resolving certain other questions. This second family of problems we shall term "technical homographies" which, while not being a case of homographies properly so-called, can, in practice, be submitted to the same methods of treatment.

For example, nouns in French are not declined, nor are nouns in English, but on the other hand German nouns are declined in four cases. A French noun can be thought of as a common homonymic form corresponding to a fictitious French declension, each of whose cases can be more easily put into correspondence with one of the declined German forms. If it is a question of translating from French into English, the pretext of homography loses all point, since the English form is not declined either. Nevertheless, it can be convenient to make use within the machine of these pseudo declensions and to associate with each noun a characteristic number, defining its function: N1 for a subject, N4 for the complementary object, etc. The usual methods of treating homographies enable us to choose between these different hypotheses: all will be tried, and we shall retain those which do not lead to any contradiction. On the other hand certain true homographies would be more easily resolved thanks to this further precision which the distinction between N1, N2, etc., carry in their context.

Figure 1.

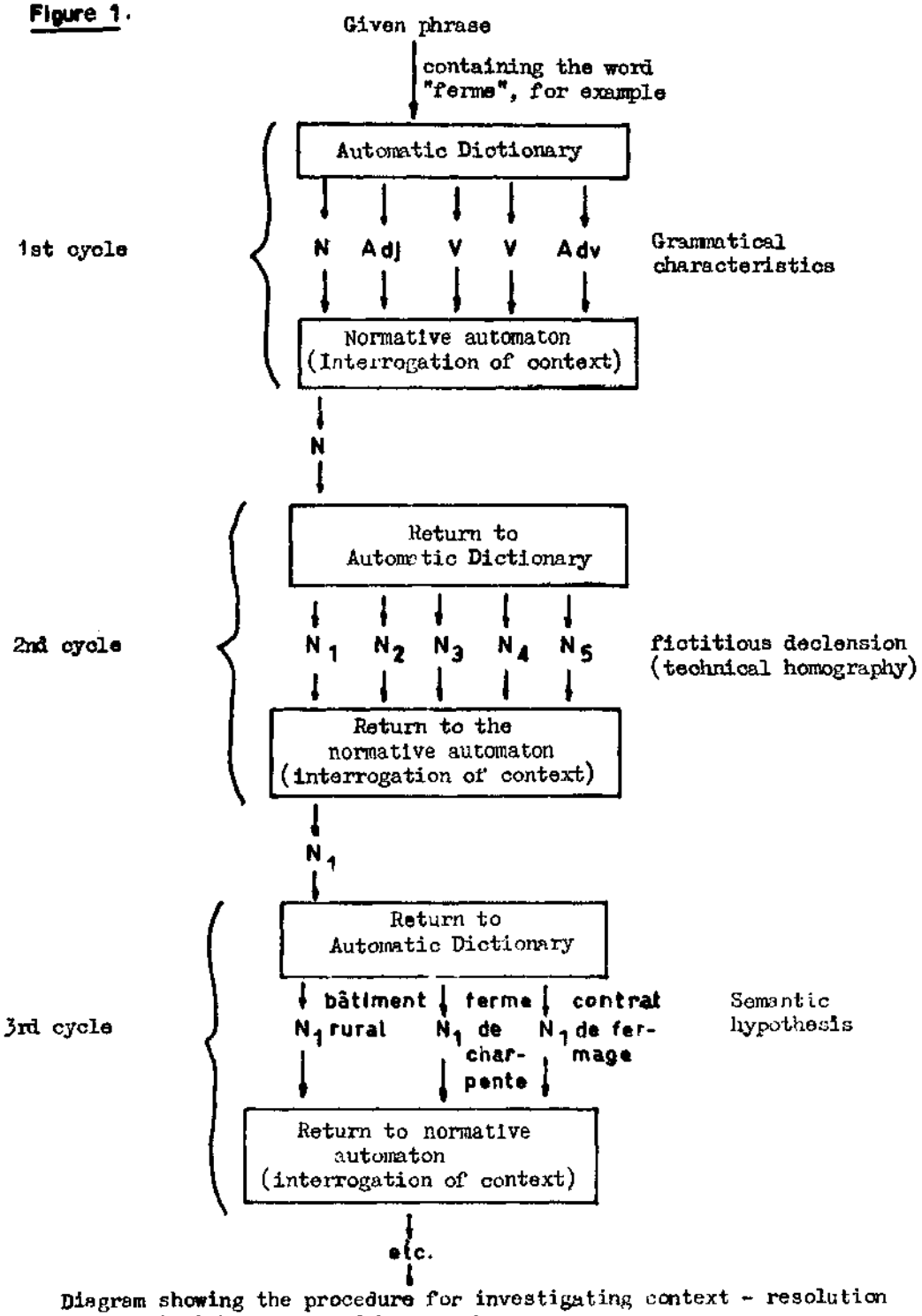


Diagram showing the procedure for investigating context - resolution of technical homonymy and homography

We see that by this expedient practically all the problems which can be posed, in the course of analysis, can be reduced to that of technical homography. To every question which can be posed concerning the words of the context, it is generally possible to give a list of answers; these different answers will be tried, one by one, and confronted with all the hypotheses which it is possible to formulate on the context of the word considered. The techniques described by the Harvard University Computation Laboratory (2) in their report NSF 4 give a good example of the procedure of interrogating a text; they lead us to associate with each word a sign characterising its essence, that is to say the nature of the role which it plays in the sentence. In the list of essences we find: subject of verb, object of verb, etc.

Nothing prevents us from pushing this interrogation further in such a way as to associate with each word not only a sign indicating its function but even a complementary sign indicating with what other words this function can be made to directly correspond; a noun will not be merely said to be a subject, it will be possible to say where its verb may be found. The parenthesis notation used in the method of context, a method of which we shall speak later, enables us to push thus far the exploitation of "technical homographies". The principle is always the same: to attempt all possible cases, that is to say all possible configurations of the parentheses and to eliminate, among these hypotheses, those which are in contradiction with the context. This principle inspires also the programmes of the Italian school (3).

An automaton capable of judgment and of accepting or refusing combinations of grammatic or semantic hypotheses can resolve all the problems of automatic analysis. The important thing is to construct a similar, normative automaton.

C) THE NORMATIVE AUTOMATON, FINITE AUTOMATON

If the role of the normative automation has been defined above with much detail and care, it is because the notion thus introduced enables us to isolate everything which, in a program for automatic translation, implies either directly or indirectly a knowledge of the input language. This knowledge is to be found, in part, in the normal automaton and, on the other hand, in the dictionary associated with it, which enables it to be fed with hypotheses concerning the given sentence.

Why is it important to localize the elements which use information of this kind? For the reason that the machines constructed by man, and translation machines likewise, can only be finite automata. But linguistic facts, in practice, constitute an infinite set. At least we are agreed to consider it as an infinite set, that is to say we never make use of the argument that, assuming we will never see, namely, a sentence of more than 1 million words for example, the number of possible sentences is finite and thus also is the language. A translating machine ought not to contain a list of sentences and translations but on the contrary make use of the fact that language has a structure and this structure enables it to describe an infinite number of events, with a finite number of rules. All the difficulty of automatic translation is here. Thus

the elements where the linguistic information is localised, ought to be examined with particular suspicion; it is these that risk being thought of as infinite automata. Imagine the scandal which would result if, having mobilized numerous teams of grammarians and mathematicians for the formation of a list, it were shown after several years of compilation work, that the projected list was, by reason of its very conception, an infinite list.

It is easy to remark that a list is, in general, constructed as a function of the procedure of look-up which ought to be applied to it and, on the other hand, that this constraint ought to impose on the list a certain level of redundancy. If the address system chosen is particularly unsuitable this redundancy level will become very high.

As the compilation of the list which assembles all the information concerning a natural language proceeds, it is necessary to supervise the level of redundancy of the list thus being made.

The most trivial case of a list with practically infinite redundancy would be constituted by a list having phrases of the source language, together with their translation. It is clear that such a list would be very easy to consult (apart of course from its size and its dimensions) and that it would give good translations, but its construction would be endless. The test of redundancy level would be very easily applied in this case. It is easily seen that any grammatical rule whatever would have its application there in a practically infinite number of entries, while the absence of redundancy supposes entries which are as independent as possible.

If we adopt the scheme of a normative automaton, we see that the information concerning the input language will be contained partly in the dictionary, as a result of which we make hypotheses on the nature of the words in the given sentence, and partly in the normative automaton which will give, in principle, a series of relations (semantic, grammatical) having to be satisfied by the combinations of hypotheses made on the nature of the words.

If the number of categories used is finite, the dictionary is finite also since the entries comprise simply, for each word, the different possible categories and their corresponding translations. This dictionary will not indeed be very much "larger", in terms of the information it contains, than certain dictionaries used by human beings.

If an element risks becoming infinite, it will certainly be the normative automaton, since it groups all the information concerning the language.

It is of interest to reunite in a list the ensemble of information furnished to this automaton, and to measure, if this is possible, in the course of its construction, the redundancy of this list. At least it will be possible to test where a particular item of information, taken at random, does not influence a practically infinite number of entries.

On the basis of the notion of non von Neuman languages (4), (5), recently brought to the fore by L. Lombardi, it is possible to put forward the fundamentals of a theory of list look-up and of list redundancy. Further publications will develop this point in detail. We here trace the main outlines.

II

LINEAR ADDRESSING AND THE PARADOX OF CATALOGUES WITH INFINITE REDUNDANCY

A) LINEAR ADDRESSING

IT is not possible to understand the normative automaton as conceived at Euratom without referring to a certain context of problems. Our sub-routines, which already simulate, in part, on a computer the work of a normal automaton, allow us to relinquish, at a certain level, those procedures of addressing which we shall call "procedures of fixed linear addressing". Let us explain why this is so.

Language is presented to us in the form of a linear sequence of signs, a linear succession of words. The most natural manner, at least at first sight, of characterizing a word in a sentence, is to indicate the rank, which it occupies in this linear sequence. The most natural manner, at least at first sight, of characterizing a set of words (and, in particular, those sets which are called "contexts"), is to indicate the ranks which they occupy in this linear sequence. Certain variants of these procedures will permit us to introduce considerations regarding the nature of the words; for example, we can characterize a noun as being the first noun which one can meet after the nth word; but the idea is basically the same.

Since, in order to decide on homonymy, we have to make appeal to the context of each ambiguous word, it is necessary to construct something which resembles a list of contexts. It is natural, before commencing to forma list, to choose the procedure of addressing which would permit it to be looked up, and the simplest choice, at first sight, is that which characterizes a context by the simultaneous fact, on the one hand, of the grammatical or semantic nature of the words which it contains, and, on the other, of their linear addresses relative to an ambiguous word of which they form the context. We do not insist on the objection, which is not very serious, which can be made out of the case where the context itself contains ambiguous words. It is necessary to study all the corresponding configurations which are finite in number and which are relatively not very numerous. The idea of constructing such a catalogue may appear at first sight reasonable and simple. Also it is convenient to recall, though it is well known to everybody, that the idea of constructing such a list, can give rise to grave misconceptions. The work to which it leads is by no means simple, and perhaps, not even reasonable.

It is not simple, if we take account of the difficulty of constructing lists of contexts, especially if they have as their aim the collection of those contexts which are never encountered. An example: an article never precedes, in French, a verb in a personal mood. Rules such as this resemble less and less the rules which are met with in current grammars. We can even see that lists with linear addressing have an appearance which is very forbidding, thus it is to be feared that the work involved in their construction will be found tedious by the grammarians of literary form. The choice of linear addressing runs the risk of depriving us of our best source of information and of those natural allies who are the thousands of linguistic specialists to be found throughout the world

But, in fact, lists which have linear addressing have another inconvenience and that on quite a grand scale: it can be easily shown that such lists, in order to be complete, ought to contain an almost infinite number of lines. Although the demonstration of this is classic, we shall repeat it in the next section, (B).

What remedies can be produced for such a situation? To our knowledge, there are two principal ways of avoiding the obstacle.

1) One consists in maintaining, despite everything, the system of linear addressing, since it appears so natural, and of constructing the list, despite everything, in an implicit form, that is to say without writing into it all the information. The subterfuge consists in this: one describes a finite corpus of configurations, that of sentences which are said to be simple; one defines, on the other hand, a procedure of generation capable, starting from this corpus, of reconstituting the whole list. (Such a procedure of generation is usually described under the name "set of linguistic transformations"). We do not naturally derive the list itself, this is not necessary, since the corpus of simple sentences and the set of linguistic transformations furnish us with equivalent information. The principle of linguistic transformation has been clearly defined by Z.Harris (6) and N.Chomsky (7). The lists remain to be constructed.

2) The other method consists in abandoning the system of linear addressing considered as a principal procedure of addressing used by the normative automaton. Linear addressing can, of course, be conserved in the auxiliary elements or prefilters destined to smooth out the work, but incapable of completely resolving the question of homonyms.

The implications corresponding to this second method of attack on linguistic problems have not, to our knowledge, ever been clearly defined, although several teams seem to be actually working towards this end. We shall show, further on, why there exists indeed a second method and why this second method of approach is characterized by discovering parameters which are said to be "intrinsic". Mathematicians are well aware that while there exists for solving a given problem a family of methods which relies on transformations, there corresponds to it in general another family of methods of solution which are said to be methods of intrinsic addressing.

But it is necessary to explain straight away why catalogues of context, edited in terms of linear addressing, may, if we do not take care, have an infinite number of lines.

B) THE PARADOX OF LISTS WITH INFINITE REDUNDANCY

The paradox is usually exposed in two steps; we show at first that the list of the normative automaton ought to contain all the grammatic and semantic rules; we then show that the inscription of a single one of these rules in a list addressed linearly occupies a practically infinite number of lines.

Amongst all the problems which are posed in the course of the step "analysis of text", it is that of homonymy (or of technical homographs)

which presents the greatest difficulty and whose solution demands the greatest amount of information. All the norms of the language, whether it is a question of grammatic or semantic rules, ought to be brought to bear on such questions.

Let us take for example grammatic rules such as, in French, the agreement in person and number of the subject with its verb; the agreement in gender and number of an adjective with its noun. It is easy to construct cases of homography, such that only the rules in question will enable us to solve these homographies. Let us cite, for example, the sentence "l'obstination de ces hommes braves la tourmente", where the agreement between the subject and verb prohibits the word "braves" from being a verb in the second person singular; let us now take the second sentence "l'obstination de ces hommes brave la tourmente", where the agreement of adjective-noun prohibits "brave" being an adjective. In the same way, once any semantic rule whatever is given, (such as that of verbs said to be "of an animate nature" having a preference for subjects which are animate beings) (8) we can construct numerous cases of homography, the solution of which, while imperfect, implies at least taking into consideration the rule in question (example: "Les pages rient", "Le facteur pleure").

As it is called upon to solve with the maximum possible efficiency the problem of homonyms, the normal automaton ought to assemble the maximum of information concerning the input language and semantics in general. This is a question of an enormous amount of information. Even in making the favourable hypothesis that they will be assembled in a catalogue of very weak redundancy we can be certain that this catalogue or list would have very formidable dimensions.

For it can be easily shown that in such a catalogue, drawn up in accordance with linear addressing and supposed complete, the item devoted to a single rule will occupy already an almost infinite number of lines, since the redundancy of this type of list is great. The example of a grammatical rule of agreement of subject and verb is particularly noteworthy. Suppose that we desire to make use of this rule in order to eliminate eventually the hypothesis that a word of rank x is a verb, for example, in the third person singular. It is necessary to verify whether the presumed subject is a singular noun perhaps, or a pronoun in the third person singular or, indeed, to find out what is the subject at all. We know that it may, perhaps, be just as well after the verb as before it. It will be necessary to envisage successively the cases where the subject has rank x + 1, a rank x + 2, etc. then the cases where the subject has rank x - 1, x - 2, x - 3, etc. and these cases are unfortunately very, very numerous. With regard to each of them, it will be necessary to note all the grammatical configurations which can exist between the subject (of rank x - n) and the supposed verb (of rank x), and these configurations are unfortunately very varied.

It is necessary, as well, to distinguish, amongst the different possible subjects, the pronouns from the nouns, etc., because different intermediate contexts will apply to them. All together it will require a practically infinite number of lines in order to write the rule of agreement of subject and verb considered in one of its applications. A fortiori also, the entire list which ought to contain all the rules, will be practically infinite.

Let us analyse the operations by which redundancy is introduced. Let us consider the question of expressing a single item of information, for example, the permissible categories for the subject when the verb is supposed to be given. In order to do this we require the entry of a number of almost infinite lines, with, at each line, this information written once, whence we have a first form of infinite redundancy. Besides, the clearest part of the work has consisted in the description of a configuration susceptible of being found between the subject and the verb, that is to say, the linguistic fact setting out not merely one, but all the rules of the grammar, whence we have a second form of redundancy, which each entry contains, of the information which figures in all others, generally in a different form, and which does not facilitate matters.

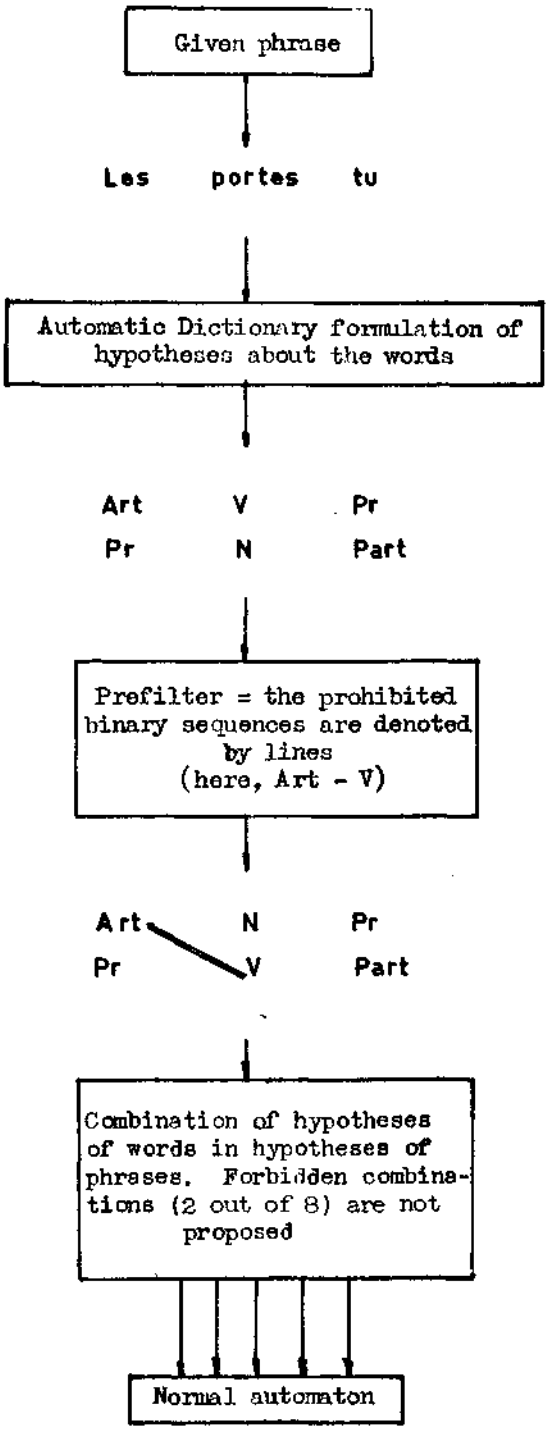
The redundancy of the list is, in general, bound up with the procedure of consulting it, with the view to which each article of this list was conceived. It happens that we voluntarily increase the redundancy of the list in order to facilitate its look up. Taken together the annual telephone directories of a town constitute an example of a list which is very redundant, since it is a question of the same three works repeated thousands of times thanks to the printing-works; not only does the information figure there often under the same form but indeed the existence of three kinds of lists; alphabetic, then ordered according to streets, then ordered according to professions, shows that the information appears there under several different forms. This redundancy is useful and nobody would think of replacing the set of these works by a single yearly non-redundant list. On the contrary, in the case of a list of linguistic norms, it appears that linear addressing imposes a redundancy level which is very excessive. When the principal obstacle to the formation of the list resides in the amount of information to be assembled and in the incumbrance which results from it, we ought above everything to avoid increasing the difficulty by imposing on the redundancy of this list a level which is practically infinite.

C) THE PREFILTERS

Before returning to the two principles which ought to govern our resolution of the paradox of lists, it would be convenient to say a few words on certain subroutines being currently made throughout the world, which do not appeal either to one or the other and for which the lists were conceived as if the problem had been simply ignored. It is usually a question of small subroutines and lists which are very limited. It goes without saying that it is not, in general, as a result of naivete that the authors of these subroutines have engaged in such work. If we were to limit the employment of such procedures to a role which is purely auxiliary, and which would consist in reducing the amount of work which the normative automaton has to do in treating certain problems of homonymy chosen amongst the most simple and most frequent, such subroutines are useful and their authors publish them (9). Though having abandoned the employment of linear addressing, in the case of the normal automaton itself, several members of our team are actually working on the

Figure 2

Role of Prefilters



construction of a prefilter by reassembling subroutines of this sort. If it is interposed between the automatic dictionary and the normal automaton, the prefilter will permit the saving of machine time by eliminating quickly certain combinations of hypotheses. Redundancy enables us to save time and it is necessary to seek a compromise.

Certain of the rules described by O.Kulagina in an article cited above can be very valuable, such as, for example, those in Appendix 1-31 relative to forbidden sequences of article-verb, or even preposition-verb.

But, at least, while it is not a question of experiments carried out on a corpus of texts which is very limited, there can be no question of treating all homonyms by the procedures of the prefilter method, without falling into the paradox of lists with infinite redundancy. Consequently, it is necessary to discuss other procedures.

D) ON THE CONSTRUCTIONS OF LISTS WITH MINIMUM REDUNDANCY

It is necessary to record the role of redundancy in the paradox of infinite lists, in order to be in a position to formulate clearly the necessity for a catalogue with minimum redundancy and to explain why there is a method other than that which makes appeal to linguistic transformations.

If we choose at first a system of addressing as is done, for example, by Z. Harris, N. Chomsky, Solomonoff (10), and if it is just a question of linear addressing which introduces the redundancy about which we know, it is necessary, as a result, to fold the list an infinite number of times on itself in order to express an item of information with the aid of "simple sentences" and of rules of "linguistic transformations".

But if we grant the simplicity and conciseness of lists without initially fixing a condition relative to the addressing, the catalogues of lists will be easy to establish. They need not make mention of all the properties of sentences but only of those which are characteristic of a sentence well constructed, those of which Chomsky says that they have invariant properties in the linguistic transformations. In order to express each property, we chose the address parameter which appears the most simple, that is to say, its intrinsic parameter. The list which results will thus have maximum simplicity, but its look-up will introduce no longer one linear parameter nor even one unique parameter but, in fact, a whole family of parameters. It will be necessary to lose in machine time a little of the advantage gained by the lists but the first is less rare and eventually more economical than the second. We can thus formulate in this survey our requirements with regard to the list to be constituted for the normal automaton;

1. The expression of each rule will be made without redundancy (or at least with the minimum redundancy). Practically each rule will be enunciated in a single line and the rule of subject and verb, in particular, will be announced in a single line.

2. Each rule will figure in the catalogue once and once only, each rule will constitute an entry and the entries will be practically independent one of the other, which enables us to formulate just the number of necessary rules.

3. Each rule will be expressed in a language which is very near to ordinary language and has the aspect of rules which will be found in grammars of the current type. It is a matter of observational rules and not operational ones. The catalogue will look like an ordinary grammar but very complete and provided with numerous semantic rules.

It is only when we respect such requirements that we can hope to receive the efficacious help of specialists of language of literary formation and, thanks to their help, to construct, in an acceptable time, a list which is indispensable to the normative automaton. When the requirements of rigour and efficiency are both satisfied we shall be then free to endow the programs with additional lists which will have redundancy as a unique function. This redundancy can save machine time but it is necessary to proportion it in such a manner as to get its optimum level. Such additional lists exist in the program of conflicts. In order to proportion the level of redundancy, it is necessary, by storing them in different places to separate the principal list (characterized by its minimum redundancy and its aptitude for solving simply a set of problems), on the one hand, and the additional lists of redundancy (which contain only information which already figures in the principal list, or of combinations of such items of information and which play a purely optional role, in such a manner that their suppression or their reduction would not have the consequence of stopping the operations but just a simple relaxation). In order to understand how a certain proportioned level of redundancy can save time we refer to the example of annual telephone directories given in paragraph B of part II.

But let us return to the problem of the construction of the principal list with minimum redundancy.

III

INTRINSIC ADDRESSING

A) THE PRECEDENT OF ANALYTIC GEOMETRY: ANALOGY OF PROBLEMS

IT is not a matter of indifference to recall that, long before linguistics, analytic geometry found itself confronted with certain difficult questions of addressing bound up with problems of transformation. The "transformational" method of attack was, in general, explored first (as in linguistics). An additional reflection is necessary to discover that when one is envisaging methods of transformational solution there generally exist also methods of intrinsic solution, more involved but more elegant.

Why precisely do we mention analytic geometry? For this reason, that analytic geometry is placed with respect to pure geometry in the same situation that automatic translation is placed with respect to pure linguistics. We would like to recall, in effect, that analytic geometry operates on three categories of mathematical entities, namely:

1) Geometric entities, properly so called (lines, planes, relations, etc.)

2) The addresses of these geometric entitles.
3) The parameters as a function of which these addresses are calculated.

The great novelty introduced by the use of computers in the problems of language resides just in the mental grasp of this difference which exists between an item of information and the address of this information. Automatic translation also operates on three categories of entitles:

1) Linguistic entitles (words, grammatical rules).
2) The machine addresses of these linguistic entities (the address of a word, the address of a grammatical rule).
3) Parameters as a function of which these addresses are calculated.

Computers also oblige us to introduce into linguistics the coordinate axes which Descartes introduced into analytical geometry, for in a machine an item of information cannot exist without an address; the address has an existence distinct from the information, the address can become an object of calculation, the address supposes, explicitly or implicitly, the choice of a parameter.

Indeed the best method of introducing the notion of intrinsic addressing into linguistics seems to us to be as follows: we shall describe in detail an example of the use of intrinsic equations in analytic geomtry, we shall show the analogy with linguistic problems, then with each of the operations which have led the geometers towards intrinsic equations, we shall associate its linguistic image, that is to say, a step of the procedure which leads to intrinsic addressing in linguistics.

The example chosen will be that of the intrinsic equations $R = f(s)$ which give the radius of curvature of a plain curve as a function of its curvilinear abscissa. Let us recall what it consists of: it is a fact of current human experience that the human eye identifies as similar two figures of a drawing, one of which does not differ from the other, except in its displacement (accompanied or not by a change of scale). Let us denote such curves as follows:

Our eye recognises, in these three figures, the manifestations of an unique shape, that of an "eight with a hump in the front", a form or shape which is felt to be altogether different from that of a seven or that of an interrogation mark.

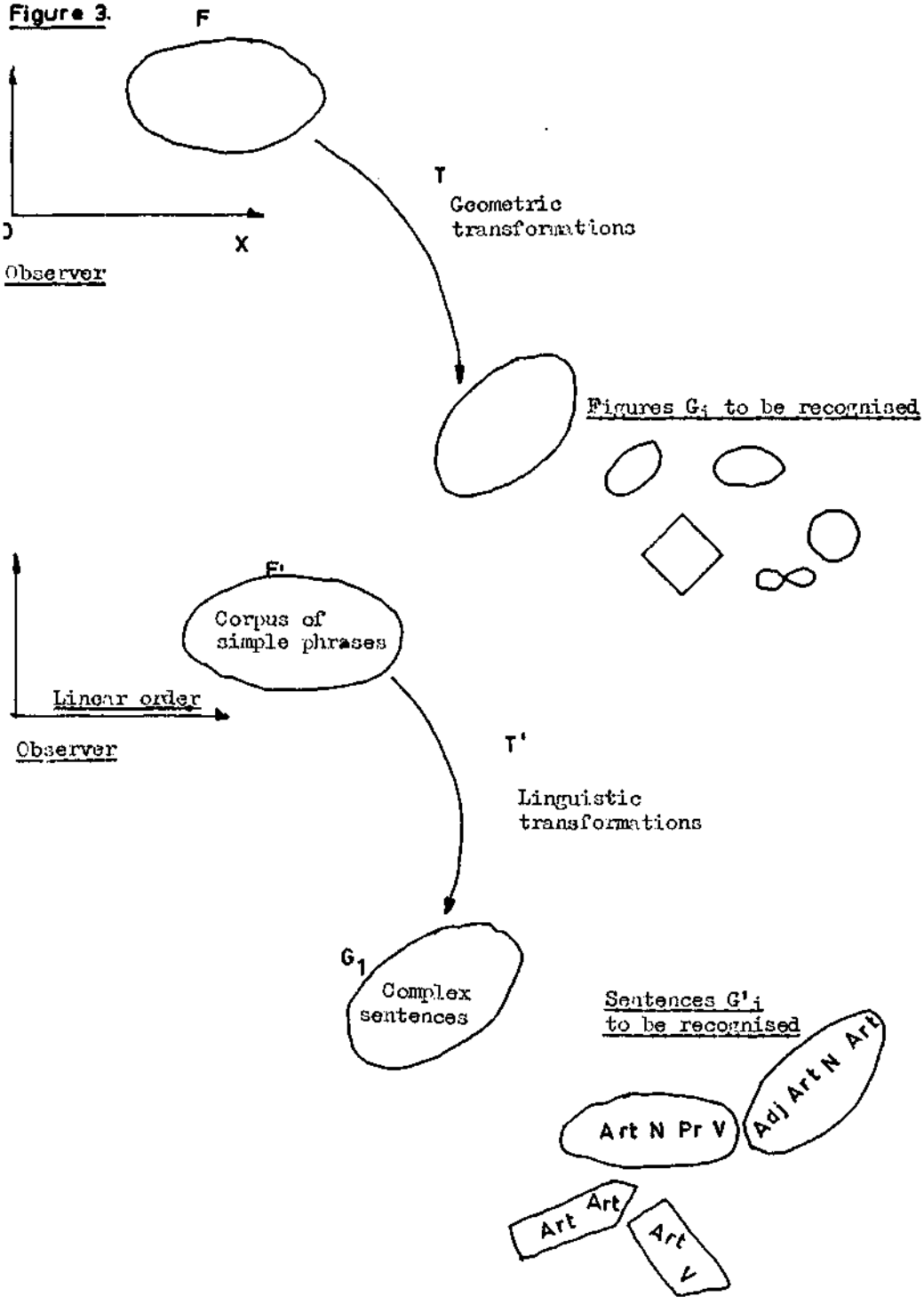
The problem of analytic geometry can be put as follows: being given a certain shape defined, for example, by a plane figure F of cartesian equation $f(x,y) = 0$, being given, on the other hand, a succession of figures G_i, whose cartesian equations are written, taking the same parameter as for F, respectively $g_i(x,y) = 0$, we want to find a procedure which, applied to any one of the figures G_i, will tell us whether it has the same shape as F or not.

We shall show at once the analogy with the linguistic problem. The image of the common cartesian parameter is the linear order of the language. Just as the figures G_i are given in this cartesian parameter so also are hypotheses of the type: Art N Adj V Art N, or even Prep Art N Art N V, etc.; hypotheses which the normative automaton has for judgment are given in the linear order of the language. In the same way, as in geometry, the problem is to verify if the figures G_i have the correct shape, so also, in linguistics, it is a question of verifying if the hypotheses constituted by the grammatical sequences are correct. That is to say, can be associated with correct sentences. (We gave an example, which is very rough, of the type Art, N, Adj etc. but it is understood that each category can be made as precise as we wish: Art masc sing, N masc sing of animate being, etc. etc.) It is a question then, in both cases, of identification of form.

Let us now pass to a description of the solutions. The geometrical problem can be treated by making an appeal to the notion of transformation, that is to say by discovering the particular family of transformations which is that of displacements in the plane (with or without, as we have said at the beginning, a change of scale). The procedure consists of noting that the infinite set of transformations of F coincides with the infinite set of figures which have the same shape as F. The question of knowing whether G_i has the same shape as F is reduced to that of knowing whether G_i is part of the infinite family of figures T_{Var} (F), transformed from F by the result of a translation of the vector V, then the rotation of an angle a and then a change of scale of ratio r. The equations of the figures $T_{Var}(F)$ can be written in parametric form and we try to find a set of values of the parameters Var which enable us to identify the equation of one of them with the given term G_i.

In a similar manner the linguistic problem can be attacked in this way by making appeal to the notion of linguistic transformations. The procedure consists in finding a corpus F' of simple sentences and a family T' of linguistic transformations in such a manner that the infinite set of correct grammatical combinations coincide with the infinite set of transforms of F'. The question of knowing whether a given combination

Figure 3.


F
T
Geometric
transformations
Observer
Figures G_i to be recognised
F'
Corpus of
simple phrases
Linear order
Observer
T'
Linguistic
transformations
G_1
Complex
sentences
Sentences G'_i
to be recognised
Art N Pr V
Adj Art N Art
Art Art
Art
V

G' is correct is reduced to that of discovering whether a combination G_i' is part of the infinite family of combinations $T'_{a'b'c'etc.}(F')$ transformed from the corpus F' of simple sentences by the linguistic transformations T'. The reference systems a'b'c' etc. characterize the particular linguistic transformation or the product of the particular linguistic transformations corresponding to each correct combination. It is understandable why Chomsky (7) speaks of the necessity of defining a process capable of engendering all the correct sentences of the language. This process is the set of transformations T' applicable to the corpus F' of simple sentences.

Let us now pass to the question of the paradox of infinite lists. In geometry, it is manifested in the following way: while F and the transformations T constitute, in their entirety, a finite sum of information, it is not possible to enumerate explicitly and completely in one cartesian parameter the list of the equations of the curve $T_{Var}(F)$, because these curves are infinite in number. The list which we are forced to form will thus have a redundancy all the greater as it will contain more lines and its redundancy will tend to be infinite. The linguistic image coincides very well with the paradox which we have announced above. To the unique cartesian parameter it is necessary to seek correspondence in the linear address in linguistics. To the list of equations, expressed explicitly, there will correspond a list of correct grammatical sequences, expressed explicitly; a list which it is impossible to construct from the fact of its dimensions, and from the redundancy which it implies.

It remains now to describe the second method used by the geometers, which is called the method of intrinsic equations, and then to find the linguistic image of this second method.

B) INTRINSIC ADDRESSING, SUGGESTED BY THE PRECEDENT OF ANALYTIC GEOMETRY

In analytic geometry, the intrinsic equation of a plane curve

$$\frac{R}{L} = \frac{f(s)}{L}$$

constitutes a sort of minimum list of properties of this curve which remain invariant in a displacement accompanied or not by a change of scale. It is, if you like, a characteristic of the shape of this curve. In this equation s is the abscissa of the point under discussion, R is the radius of curvature, and L is a characteristic length measured on the curve, for example, the greatest diameter or even curvilinear length of an arc joining two singular points, or indeed even the period s if F is a periodic function, etc. If the figure F satisfies such an intrinsic equation, the set of transformations of F, and these alone, satisfy it equally well. Any figure G_i being given, it will be sufficient, to discover if it is similar to F_i, to verify whether its curvilinear abscissa and its curvature satisfy the intrinsic equation.

How do the geometers arrive at such a result? By using, for addressing the information of the list thus formed, not only a single cartesian

parameter but a whole family of such parameters, namely those constituted by the set of the tangent-normal pairs of the curves studied.

Let us compare, from the point of view of the size of the lists, the respective merits of the transformation method and the intrinsic method.

TABLE I

List of the Transformational method	List of the intrinsic method
1) Cartesian equation of the figure F. This equation defines:	Invariant properties in the transformations.
- the invariant properties in the transformation T_{Var}	
- the other properties of F	
2) Description of the transformation T_{Var}	

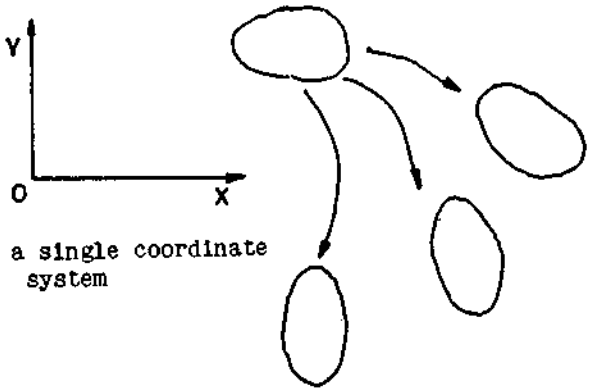
Since the technique of linguistic images has given us satisfactory results so far, we ought to be prepared to find by means of this technique, bearing in mind the theory of intrinsic equations in analytical geometry, an intrinsic method for the resolution of linguistic problems, such as those of homonymy. The necessary lists for effecting such an intrinsic method ought to contain solely the list of the linguistic properties which are invariant in the linguistic transformations. It will be permissible to use, for addressing these properties, not only a single reference system but as many different reference systems as will be necessary for expressing these properties in a simple fashion. We could, if we wished, utilize a different reference system for each property inscribed in the list. To the comparative table given above, for geometry, there will correspond the following linguistic image:

TABLE II

List of the Method of linguistic transformations	List of the method of intrinsic addressing
1°) Description of a Corpus F' of simple sentences. This description defines	Invariant properties in the linguistic transformations
- the properties of these sentences which remain invariant in the linguistic transformations	
- the other properties	
2°) Description of the linguistic transformations $T'_{a'b'c'}$ ...	

Figure 4.

Non-intrinsic addressing



a single coordinate system

An infinity of equations

$S_1 (X,Y) = 0$

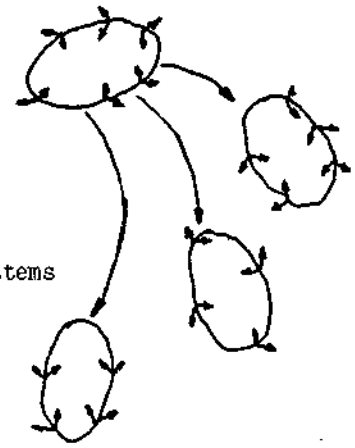
$S_2 (X,Y) = 0$

$S_3 (X,Y) = 0$

$S_4 (X,Y) = 0$

. . . . . .

Intrinsic addressing



An infinity of coordinate systems

A single equation

$f\left(\frac{R}{L}, \frac{S}{L}\right) = 0$

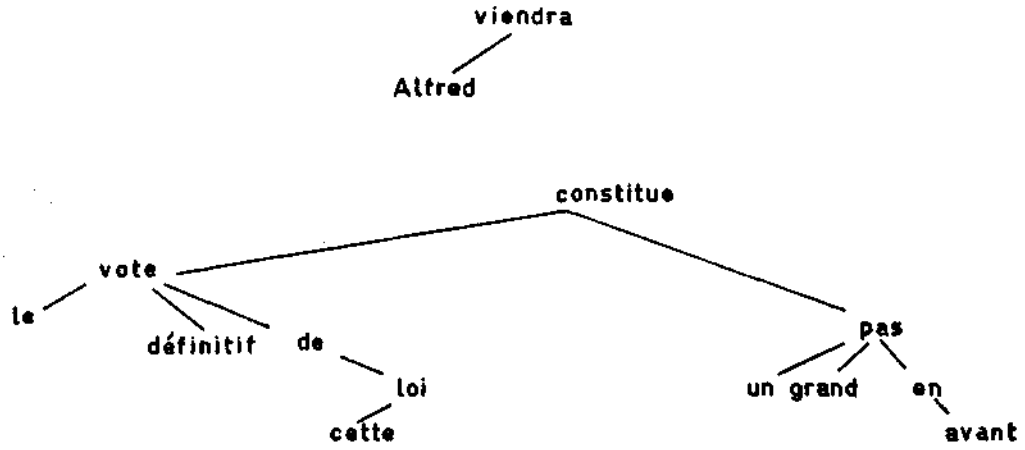
The considerable gain realized from the factor of simplification of the lists will probably be compensated by a complication of the system of look-up. But, as the principal difficulty in automatic translation resides in the formation of the complete lists, we ought to expect a very positive balance.

Let us analyse the steps which the mathematicians take when they use intrinsic methods and attempt to apply these same principles for resolving linguistic problems. The idea is to give priority to the requirements relative to the list. If we try to describe geometric properties of a figure, it can easily be seen that this description can, in general, be reduced to an enumeration of the set of local properties relative to a neighbourhood of each point of the curve considered. For, the local invariant properties are, for a curve, those which fix the local value of the curvature and the local evolution of this curve as a function of the curvilinear abscissa. The simplest parameter for expressing these properties is that constituted by the tangent-normal pair at the point considered. It is the set of these local items of information which one assembles in an equation of the form $R = h(s)$. In order to take account eventually of a change of scale, we write

$$\frac{R}{L} = \frac{f(s)}{L}$$

Thus it is the use of the best adapted local parameter for the local item of information which enables us to make up a minimum list.

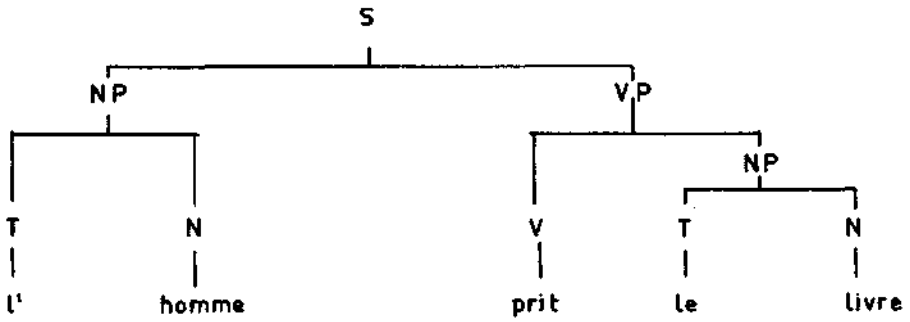
Let us apply the same principles in the case of linguistics. If we take again the example of the subject-verb agreement as an example of a property to be put in the list, there is no simpler parameter in which to express this fact than that constituted by the feature which directly links these two words; this is what has just been done in the stemmas of Tesniere (11), and in the program of Hays (12).



Whatever may be the number of words which separate them in the linear order of the sentence, the addresses relative to the subject and verb always remain the same in the diagrams given above and we cannot imagine a simpler address than the direct bond. If we use this bond as a parameter, the law of subject-verb can be written in a few words; the rules of agreement, an indication of the fact that one subject can only be the subject of a single verb and that a verb can have only one subject. This is all. The case of co-ordinate subjects has not been mentioned, because the conjunctions of co-ordination introduce general exceptions, which are indicated in another rubric of the grammar.

Let us take another example: in order to describe the agreements between a noun and its article (the rule of agreement; an article determines a single noun and a noun accepts but a single article), what better reference system can we have than a feature which directly links both these words? We are thus lead to note the respective agreements between subject and verb, and article and noun, not merely in the same reference system but according to different reference systems constituted each by a direct link, between the words concerned. This is what has been produced in the diagrams of Tesniere and Hays; and we see that such diagrams have every justification in the list of methods of intrinsic addressing. In a general way, the systems where the words appear bonded together, two by two, lends itself very well to the addressing of corresponding grammatical relations which are of a binary nature, between the words.

On the other hand, for the expression of grammatical relations which bind together not merely two words, but two groups of words, other sets of parameters are more suitable, such as, the following set:



This second type of diagram is described by Bar Hillel (13), Yngve (14), Chomsky (7). It is also very suitable for the addressing of grammatical rules which are called "traditional"; because to express the agreements between subject and predicate, what better reference system than a feature which directly links both these groups of words? In order to express agreement between the verb group and the complementary group of the object, what better reference system than a feature which directly links them? And so on. We see that the "diagram of derivation" above constitutes not merely a unique reference system but, indeed, as has been said above, a set of reference systems each one of which has a certain local usefulness.

There are many other diagrams which are favourable for linear addressing of either grammar or semantics. (cf. Bibliography) This diversity need not astonish us. The choice of reference systems is always

relatively optional in analytic geometry and there is no reason why this freedom and this partial arbitrariness should not exist in linguistics for the choice of intrinsic reference systems. Without doubt some of them are more suitable than others, but in the neighbourhood of the optimum these differences are very small.

The choice of a reference system or a diagram grouping a family of reference systems need not prevent us, in general, from using simultaneously many others. While two reference systems are respectively suitable for the expression of two independent laws, nothing prohibits us from using them together. If two families of reference systems can be used respectively for two groups of laws which are not independent, it is necessary to verify certain conditions of compatibility (15). If these latter are satisfied, the simultaneous use of the two families of reference systems is possible. For example, the program of conflicts (16) operates at the same time on the diagram of Tesniere and that of Bar Hillel, which enables us to give addresses, on the one hand, to the grammatical laws between words and, on the other hand, to those which bind the groups of words to each other.

Thanks to all these facilities and thanks to the fact that laws of assembly resemble those of ordinary grammars, we can hope to compile lists rapidly.

Let us now show how such a list can be used, that is to say how it can resolve the initial problem mentioned above: being given a sequence G_i of grammatical categories (for example: Art, N. V, etc.) to say whether G_i' can be associated with a sentence of the language or, in other terms, to say whether G_i' possesses the set of properties which characterize correct sentences. We recall, in fact, that it is to this question precisely that the normative automaton ought to be able to give an answer in every case.

Each one is free to invent methods of look-up of intrinsic lists; if we choose a well-adapted method, the operations are much more rapid. But to show that this look-up is, in fact, possible and enables us to obtain the desired result, we will describe an absolutely general procedure as follows: the sequence of grammatical categories G_i' which is given comprises a finite number n of "words", the same number as the sentence to be analysed. If we bond these n words by features of every possible kind we can derive the set of systems of addressing which can be associated with sentences of n words. These systems are finite in number and computers can calculate them quickly. Each one of these systems binds the n terms of the sequence G_i' and it is possible to verify if a particular assembly is, or is not, the intrinsic system of a correct sentence. If the answer is "yes", each bond of the system ought to lead to a rule which really exists in the list (example: for the bond Art - N we search to see if there exists a rule Art-N) and this rule ought to be respected in all its details (the agreement of two words, etc.). In order that a system be rejected without further checking, it is sufficient that one simple bond is recognized as being wrong, either because it does not lead to any rule (for example: Art - V) or because it leads to a rule without obeying it. If eventually there is no system accepted, the sequence G_i' proposed is not a correct sentence. If one or more systems are accepted, the sequence G_i' is correct with respectively one or more possibilities of interpretation.

This general procedure which is very inelegant can be considerably speeded up without losing any of its rigour if we match the systems with the list as we make them up. If, for example, the first bond proposed

is a bond article-verb, absent from the list, not only is the bond destroyed but, in fact, all the virtual systems which, if they had been constructed, would have contained this bond, are eliminated beforehand; that is to say, eliminated before they are even constructed; none of them are even proposed.

Figure 5.

Formation of Combinations of Hypotheses

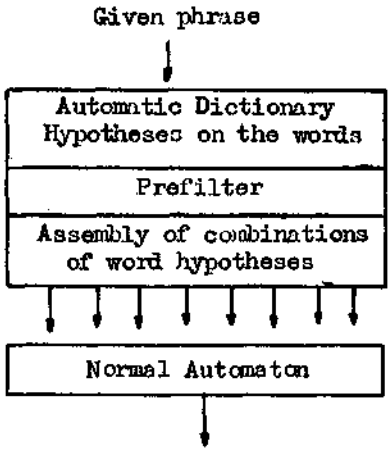


Figure 6

Detail of the treatment of a combination of hypotheses in the Normal Automaton
General Method - slow kind

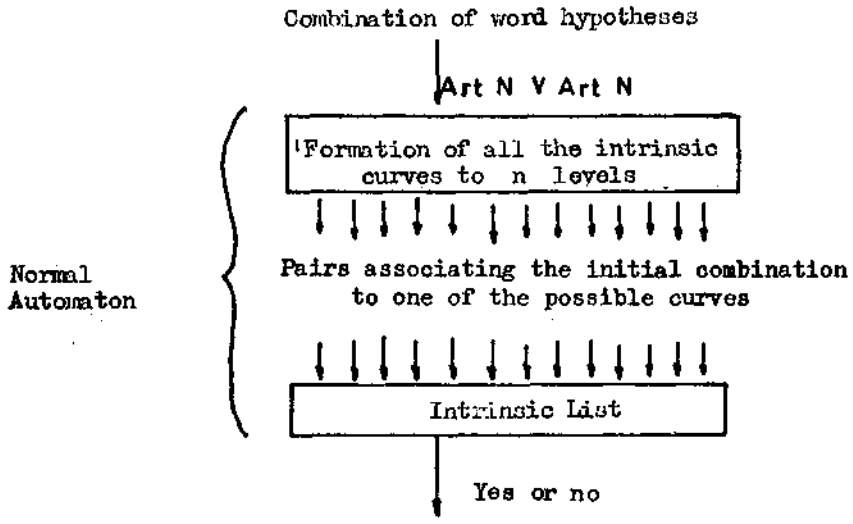
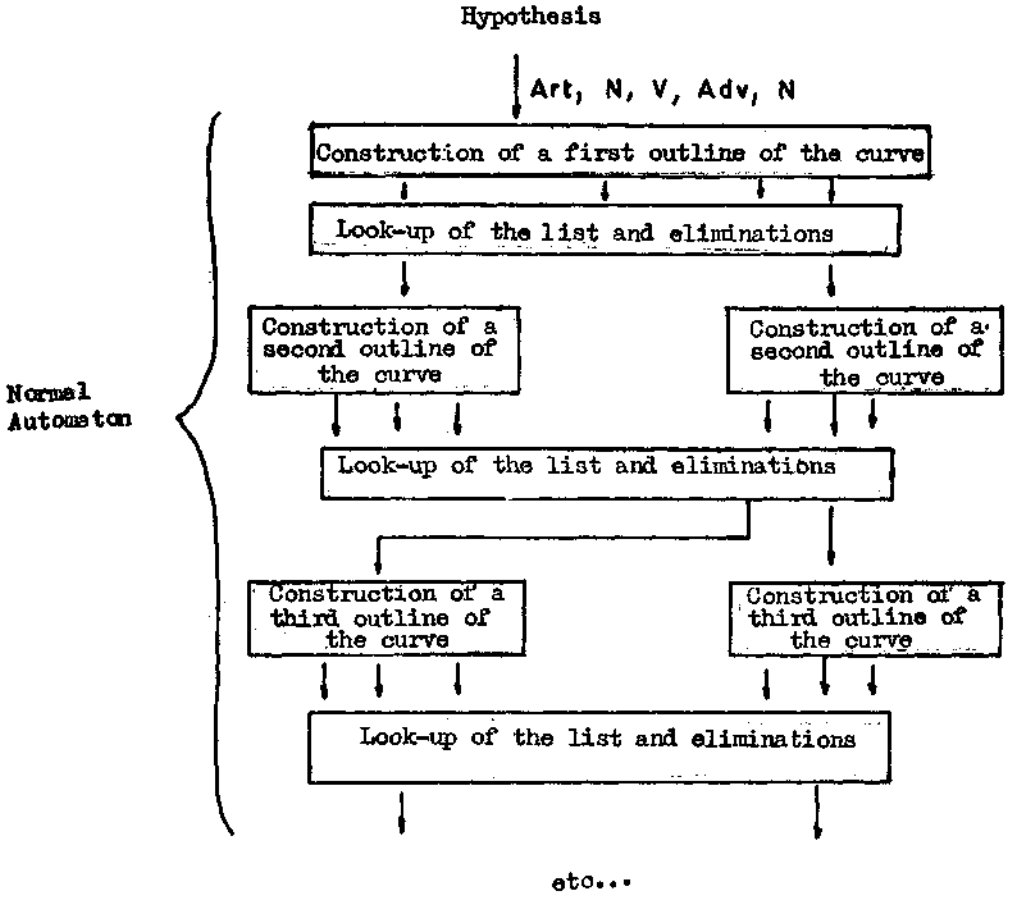


Figure 7 General Method - speeded-up kind (elimination, in packets, of virtual curves)



Thus instead of eliminating the systems one by one, we eliminate the virtual systems in large packets and the operation proceeds very much more rapidly.

Independently of the numerous improvements which can be introduced, the very existence of a general method is of great importance theoretically. It shows that, for the resolution of homographies, the intrinsic methods form a means of attack at least as legitimate as those which have inspired transformational analysis. Thus, the analogy with certain problems of analytic geometry has suggested the existence of an entire family of methods which can be used for automatic translation.

C) THE RECURSIVE CHARACTER OF THE INTRINSIC PROCEDURE

It would not be very reasonable to attempt to resolve, at a single swoop, the problem of the choice of the intrinsic addressing method for the treatment of homographies in a given language; there would be much difficulty in reconciling the two conditions as follows:

1) The form which it would be convenient to give to the entries in the list is conditioned by the look-up procedure which shall be applied to it. Definitive editing of the list thus ought to follow and not to precede the choice of method of addressing.

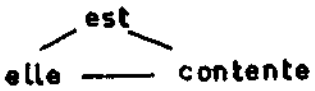
2) But, on the contrary, the choice is conditioned by the knowledge, at least in essentials of the information to insert in the list or at least by the knowledge of certain characteristics of this information.

We could, to be sure, imagine ourselves assembling, at first, an enormous quantity of information which would be considered as sufficient for resolving the problem of homographies; the method of addressing will then be chosen; the form of entry in the list will then be determined, and it will be possible to rewrite the information.

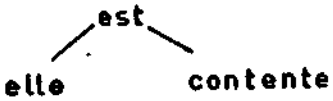
But, in fact, the moment we choose a certain kind of intrinsic method we can never be certain of taking this decision with the fullest knowledge of our reasons why. Thus it is preferable to foresee the eventuality of one or several imperfect choices in such a way as to be able to utilize the consequences of an omission or of the absence of an important item of information at the time of the first choice. We proceed by successive approximations.

A certain flexibility in the conduct of the operations results from the fact that the addressing operation is a transitive one: if the position of A is known with respect to that of B and that of B with respect to that of C then the position of A is known with respect of that of C. This property enables us to refrain from associating with certain grammatical or semantic agreements a direct addressing bond, if we wish, on condition that we utilize other bonds as supports.

Example



Sestier's notation (17)



Tesniere-Hays' notation (11) (12)

If this procedure is abused, we are in danger of having redundancy invade the list again. In the case of indirect addressing, in fact, it is necessary to name the intermediate words, for we must bear in mind the injurious consequences which can result, in linear addressing, from this necessity of naming the intermediaries. When we find ourselves in the presence of a rule which cannot be addressed directly, it is easy to see if its indirect addressing can be done cheaply or not. If the answer is yes, we adopt this as our procedure. If not, we have to put the rule on one side and wait for a later stage when, thanks to the strengthening of the system, its addressing can be done at a reasonable price.

In practice, we commence by choosing a first system which is very simple, for example, that of linear addressing, which is suitable for relations over a short distance. On this first system we superpose, at a second stage, a diagram which is suitable for the addressing of bonds of medium distance, such as those of Tesniere or Hays or Sestier, or even the correlational system of Ceccato, etc. The majority of the rules can be expressed economically and, from this stage, it is possible to achieve good programs. Perfection however is not yet attained. In order to inscribe certain rules in the list, another stage is necessary as is also the construction of a system covering a great distance. And so on; we can approach as nearly as we like to perfection without, however, modifying the parts of the list which have already been edited.

These procedures introduce into the list a natural stratification (short distance, medium distance, etc.) which can be taken advantage of in the program of look-up; information about the linear order will facilitate the construction of the system covering medium distance and information about this latter (even under the form of several hypotheses) will facilitate, in its turn, the formation of the system covering a long distance and so on.

CONCLUSION

It will seem that all that has been said is but a question of commonsense; it is necessary to agree on a methodology before discussing, with figures to hand, any economies which any particular procedure might bring about. It is only with difficulty that a program can be discussed, apart from its context. Our programs, of which the most recent and most complete description is contained in report CETIS No.23, and where an account is given of experiments conducted in December 1960, only make sense in terms of the principles which have just been discussed.

REFERENCES

(1) RABIN, M.O. SCOTT D. "Finite Automata and Their Decision Problems", *IBM Journal,* 1959.

(2) BOSSERT W
GUILIANO V.
GRANT, S
"Automatic Syntactic Analysis of English" The Computation Laboratory, Harvard University, Cambridge, Massachusetts, Report NSF 4.

(3) ALBANI E
"Construction of the correlational net by means of digital computer". International Conference for Standards on a Common Language for Machine Searching and Translation, Cleveland, 1959.

(4) LOMBARDI L.
"Theory of Files" *Proceedings of the 1960 Eastern Joint Computer Conference.*

(5) LOMBARDI L.
"System Handling of Functional Operators". *Journal of the Association for Computing Machine,* (to appear, July 1961).

(6) HARRIS Z.
"Co-occurrence and transformation in Linguistic Structure" *Language,* No.33 1957.

HARRIS Z.
"Lingustics transformations for information retrieval", preprints of papers for the International Conference on Scientific Information, National Academy of Sciences National Research Council, Washington DC 1956

(7) CHOMSKY N.
"Syntactic Structures", Mouton and Co. S-Gravenhage, 1957.

(8) CECCATO S.
"Principles and classifications of an operational grammar for Mechanical Translation". International conference for Standards on Common Language for Machine Searching and Translation, Cleveland, 1959.

(9) MARETTI E.
"How to represent and rule correlating" International Conference for Standards on Common Language for Machine Searching and Translation, Cleveland, 1959

(9) KULAGINA 0.
"French to Russian Machine Translation", *JPRS* 1961, No. 6494, Translation of an article in the Russian Language publication *Problemy kibernetiki,* Moscow, 1960 No. 4, pages 207-257.

(10) SOLOMONOFF R. S. "A new method for discovering the grammars of phrase structure language". *Proc. ICIP,* Paris: 1959 Unesco, Paris.

(11) TESNIERE L. "Elements de syntaxe structural". Klincksieck, Paris 1959.

(12) HARPER K.E. and HAYS D. G. "The use of machines in the construction of a grammar and computer program for structural analysis". *Pro. ICIP,* Paris, 1959, Unesco, Paris.

HAYS D. G. "Basic principles and technical variations in sentence structure determination". 4th London Symposium on Information Theory 1960

(13) BAR HILLEL Y. (Heb. U.J.) "Report on the state of MT in USA and Great Britain, 1959", Technical Report No.1 prepared for US Office of Naval Research, Information System branch, Contract NONr 2578(00) NR 049-130, 1959.

(14) YNGVE V. "A model and an hypothesis for language structure". *Proc. Amer. Phil. Soc.* 1960.

(15) LECERF Y. "Une representation algebrique de la structure des phrases dans diverses langues naturelles". *Notes aux comptes rendus de l'Academie des Sciences,* 1961, 252, No. 2, p.232

(16) LECERF Y. "Programme des conflits, modele des conflits", *Bi-monthly bulletin of l'Association pour l'etude et le developpenent de la traduction automatique et de la linguistique appliquee* (Atala) No.4, (1960), and No. 2 (1960).

(17) SESTIER A. "La traduction automatique" *Ingenieurs et Techniciens,* March 1959, April 1959, May 1959, June 1959.

SESTIER A. Preface to the note CNRS No. 3 (March 1960).