STRUCTUURDEFINITIES MET BEHULP VAN EEN COMPUTER. VAN WOORD TOT MORFEEN

J. DE KOCK, Hoogleraar K.U.L.

W. BOSSAERT, Eerstaanwezend assistent

A computer is a machine for data processing which can derive the final information (output) required from a given amount of primary information (source data or input) by using a number of appropriate techniques.

This processing presents a close analogy with human perception. Man uses a code for the communication of information, the so-called natural language. When this language code and an appropriate processing procedure are supplied as data to a computer one can get the implicit information present in the language code in an explicit form in a fully automatic way. The study of data processing techniques used with a computer system can be useful for the study of man and his linguistic communication system.

All information supplied to a computer has to be complete and well defined. The system cannot use unknown information or information which has not been specified in full details. When only simple linguistic forms without any reference to semantics or etymology are available to the computer system, one can be sure that the result obtained in an automatic way will not contain a single reference to the meaning or origin of the linguistic segment examined. An examination by means of a computer is a guarantee for objectivity.

For automatic processing the information has to be in an appropriate form and it must have a limited volume. For the study of the spoken language a system is needed by which every articulated form can be represented in an unambiguous way by a finite number of symbols.

The way in which the end result can be obtained must also be described in full details in a program. In the algorithm one must make allowance for all cases that are possible in theory. One can never obtain a result by an operation for which no provision has been made or which had not been described before-

Extrait de la Revue (R.E.L.O.)
VII, 1 à 4, 1971. C.I.P.L. - Université de Liège - Tous droits réservés.

hand. When the automatic result seems to be unwarranted according to the speaker of the language, it may be assumed that the criteria used in the processing were either inappropriate or incomplete or arranged in the wrong way.

The study of a complex information system, e.g. the language code, can remain flexible if one uses a computer. The same primary information can be examined in successive stages with different criteria and the validity of a given theory can be checked by applying it to a great number of different cases.

The computer has also got its limitations. Some qualitative criteria can hardly be expressed quantitatively, or be digitalized and because of this they cannot be incorporated in the program in an accurate way. In a computer all operations take place one after another. Even if the speed for every single operation is high it may take a very long time to obtain one particular survey of a complete body of information. In such cases the person in charge of the research project has to split up the work, limit the number of information elements or simplify the algorithm and deal with fewer criteria.

The investigation discussed in this contribution is part of a more comprehensive project, viz. the study of various aspects of several languages, most of them belonging to the Romance group, by means of a computer.

This study was undertaken by a team consisting of a linguist and a mathematician specialized in computer sciences and it was executed with the computer of the Digital Computer Centre of Ghent University. This computer is an IBM 360 type 30 with a core storage system of 64 bytes and with an external storage system consisting of three magnetic disk storages, type 2311 and two magnetic tape storages, type 2401.

This contribution deals exclusively with the French language code and it has been restricted to the examination of the standardized segments of the spoken language : these comprise all inflected and conjugated forms. The international phonetic alphabet was used for the transcription; individual and syntactic phonetic differences have not been taken into account. The number of segments to be examined was limited to a very great extent : only forms occurring in the *Frequency dictionary of French words* by A. Juilland were examined in the phonetic transcription of *Warnant, Dictionnaire de la prononciation*

française. This amounted to a total number of 8782 different forms after having combined the homophonous forms. Such a list can in itself represent the vocabulary of an individual and even of a small community and consequently it can supply results that are valid for the human memory within the limits just defined.

In a series of preceding studies (1) dealing with the elements which are necessary, and in themselves sufficient, to distinguish a word from any other word of a given vocabulary, it was shown that the information volume of the phonemes is very big and even bigger than absolutely necessary. One explanation for this phenomenon reads : language is very uneconomical and irrational and the non-distinctive phonemes are completely redundant. But an other and better explanation would be : the phonemes, the smallest distinctive units do not only serve to distinguish the words from one another but in certain circumstances they can be the carriers of additional information. This will be discussed in our contribution.

The scientific research into a number of phenomena consists essentially of finding invariable elements by the comparison of these phenomena. This has been our approach. The forms of the vocabulary have been compared with one another, not with the intention to find out in what respect they differ from one another, but to find out in what respects they resemble each other.

By using lists in which all forms of the language have been arranged some of them from left to right and others from right to left one can easily find that there are phonemes and phoneme groups used in the same way in more than one form of the vocabulary. Consequently, a linguistic form can be split up into two segments : an invariable segment that has this form in common with a whole series of other forms and a variable segment which is different in this series. The form ɑbʀite can be segmented into ɑbʀit and *e*, the segment ɑbʀit being invariable in the series ɑbʀit, ɑbʀitɑ̃, ɑbʀitɑ, ɑbʀitwɑ, ɑbʀitwe, ɑbʀitwɛ, ɑbʀitwɛl, ɑbʀitwɛlmɑ̃, ɑbʀitwɛʀ, ɑbʀitjɔ̃, ɑbʀitje, ɑbʀity, ɑbʀityd, ɑbʀityʀe, etc. whereas the segment *e* alternates with the zero segment and with the segments ɑ̃, ɑ, wɑ, we, wɛ, wɛl, wɛlmɑ̃, wɛʀ, jɔ̃, je, y, yd, yʀe, yʀɛ, ɑsjɔ̃, ɛ, ɛʀ.

In the following sections we shall indicate a possible segmentation by the sign :: and for the indication of a series of alternating segments we shall use the symbol := with the meaning "can be replaced by"; we shall use the symbol / for "or" and the symbol ∪ will stand for the zero segment. So in the form ɑbʀite the segmentation ɑbʀit::e is possible because of the alternation rule e := ∪ | ɑ̃ | ɑ | wɑ | we | wɛ | wɛl | wɛlmɑ̃ | wɛʀ | jɔ̃ | je | y | yd | yʀe | yʀɛ | ɑsjɔ̃ | ɛ | ɛʀ.

By the systematic application of this criterion one can indicate a possible segmentation in any linguistic form and even several segmentations, all differing from each other. In the same form ɑbite we can also have the segmentation ɑbi::te beside ɑbit::e for, the segment ɑbi is invariable in the words ɑbi, ɑbij, ɑbijɑ̃, ɑbije, ɑbil, ɑbilte, ɑbim, ɑbit, ɑbitɑ̃, ɑbita, ɑbitwɑ, ɑbitwe, ɑbitwɛ, ɑbitwɛl, ɑbitwɛlmɑ̃, ɑbitwɛʀ, ɑbitjɔ̃, ɑbilje, ɑbity, ɑbityɑ, ɑbityʀe, ɑbityʀɛ, ɑbitasjɔ̃, ɑbitɛ, ɑbitɛʀ.

In an economical system, however, priority should be given to the longest common segment. Therefore the reasoning to arrive at the segmentation ɑbi::te must not be based upon forms with a longer invariable segment, e.g. ɑbit, ɑbitɑ̃, ɑbita, etc. In the alternation series for *te* only the segments which do not begin with *t* can be maintained : te := ∅ | j | jɑ̃ | je | jɛ | l | lte | m.

The same operation can be performed for ɑb::ite with the result ite := ɔ̃d | ɔ̃dɑ̃s | ɑ̃ | e | ɛs |...,
(a series of 50 alternating segments; none of them beginning with i), for ɑ::bite and for other forms, e.g. divɛʀsite, ynite, etc.

With this algorithm one can draw up a second alternating series even when the same possible segmentation is maintained, namely by considering the variable segment as an invariable one and by comparing the form with other linguistic forms. In ɑbit::e one can consider *e,* in its turn, as an invariable segment when one compares ɑbite with kʀee, be, sykɔ̃be, tɔ̃be, ʀtɔ̃be, ɑbe, bebe, kuʀbe, ɑbsɔʀbe, deʀɔbe, ɑbɔʀde, fɔ̃de, kɔ̃fɔ̃de, ʀepɔ̃de, gʀɔ̃de, defɑ̃de, etc. (a single series of 1034 forms in the vocabulary examined). In ɑbi::te, *te* can also be considered as an invariable with the alternating series ɑbi := fɛʀmə | puʀə | ʀʒə | ɑ̃pʀœ̃ | bɔ̃ | kɔ̃ | ʀakɔ̃ | vɔlɔ̃ | mɔ̃ | ʀmɔ̃ | syʀmɔ̃ | ɑfʀɔ̃ | ɑ̃ | ɔʀiɑ̃ | plɑ̃ | mɔ̃ | ɔgmɑ̃ | ɑlimɑ̃ | sɑ̃|...
(136 alternating elements in the vocabulary). In order not to complicate the exposition unduly these longer series will not be considered in our article. From the course of the examination it will appear that they are not useful.

By using a computer it is possible to find a great number of alternations that would probably have escaped our notice in an ordinary intuitive examination. The series abready mentioned produced unexpected results. In the series for ɑbit::e we found e := wɑ | we | wɛ | wɛl | wɛlmɑ̃ | wɛʀ | y | yɑ | yʀe | yʀɛ, and in the series for ɑbi::te the alternations one would not expect are : te ::= j | jɑ̃ | je | jɛ | l | m . Other examples are : ite ::= ɔ̃d | ɑ | ɑty | ʀi | ɔʀ | ɔʀd for ɑbi::te and ite ::= ɔ̃ | ɔ̃d | ɔ̃dəmɑ̃ | ɔ̃dœʀ | ɛs | ɛse | ɛsjɔ̃ | ɛsjɔnɛl , ɛɔe for the possible segmentation pʀɔf::ite in pʀɔfite etc.

By writing down the alternating series consistently the shortcomings of the algorithm can be detected. By successive modifications it is possible to obtain a better algorithm which in its turn can be checked by using the new lists supplied by the computer. One can understand why traditional linguistics, which had to do without the help of a computer, had decided that only semantics and etymology could eliminate the shortcomings of a simple alternation rule.

So the same linguistic form can be segmented into various ways and several alternation series are possible. This fact shows that the indication of an invariable element together with an alternation rule does not suffice to proceed to a given segmentation. The alternation rule must meet a number of additional requirements. If one compares the different alternation series with one another one can find out to what extent a given alternation is repeated. If the alternations of one series can also be found in some other series, this series has some kind of general validity. The alternation series related to ɑbit::e is more general than the series related to ɑbi::te because there are many forms ending in an *e* which alternates with elements from the alternation series related to ɑbit::e . For instance, kʀe::e because of kʀe , kʀeɑ̃ , kʀeɑ ., kʀeɑsjɔ̃ , etc. ɛ̃pɔʀt::e because of ɛ̃pɔʀt , ɛ̃pɔʀtɑ̃ , ɛ̃pɔʀtɑ , ɛ̃pɔʀtɑsjɔ̃ , etc. In the vocabulary there are 925 forms in total. The number of forms ending in *te* that alternate like the *te* in ɑbi::te amounts to 65 only, i.a. da::te because of dal and dam', the form paʀte because of paʀl , paʀjɔ̃ , paʀje , etc. The alternation series for bite belonging to the segmentation ɑ::bite is not general at all.

The number of possible segmentations with the same alternation series is a yardstick for the general relevance of the alternation rule. We call this *f*. The values for ɑbite are : f (ɑbit::e) = 925, f (ɑbi::te) = 65, f (ɑb::ite) = 11, f (ɑ::bite) = 0. So, the series belonging to the segmentation ɑbit::e is the most general one. The f values for the different possible segmentations of the form ynite are f (ynit::e) = 0, f (yni::te) = 66, f (yn::ite) = 30, f (y::nite) = 1.

From these data we can draw the conclusion that the series belonging to yn::ite is less general than the series belonging to yni::te . The alternation series of yn::ite is simple, however. If one eliminates all elements of an alternation series that are not valid in any other case, one obtains a more compact rule. The number of remaining elements *(n)* is a yardstick for the complexity of the series. The condensed rule belonging to yni::te is : te := u | k | R | ʒɔ̃ | se | sɛ | vɛR and for yn::ite it is : ite := u | jɔ̃ . The value of *n* in the case of yn::ite equals 2, but it amounts to 7 in the case of yni::te .

If these two characteristics, namely the general relevance and the simplicity, are taken into account simultaneously by applying the formula $v = f/n$, every segmentation complying with the criterion invariability-alternation can get a value *v*, which can be compared with other *v* values. In the case of ynite we obtained the following values : v (ynit::e) = 0, v (yni::te) = 66/7 = 9.4, v (yn::ite) = 30/2 = 15, v (y::nite) = 1/1 = 1.

From this we infer that the segmentation with the highest v value, yn::ite , is the most economical and consequently the most likely and preferable one. If we use this completed algorithm for the case of ɑbite the segmentation ɑbit::e appears to have the characteristics just mentioned, for, v (ɑbit::e) = 925/16 = 57.8 is much higher than the values for the other possibilities : v (ɑbi::te) = 65/6 = 10.8, v (ɑb::ite) = 11/12 = 0.9, v (ɑ::bite) = 0 (after the elemination of the redundant elements in the rule for ɑbit::e remained : e := u | ɑ̃ | ɑ | wɑ | we | ~~wɛ~~ | wɛl | jɔ̃ | je | y | yɑ | yRe | yRɛ | ɑsjɔ̃ | ɛ | ɛR , and in the case of ɑbi::te we had te := u | j | jɔ̃ | je | l | m).

After the examination of a great number of series and after the application of a scale factor that can depend on the kind of series and on the kind of language examined, there appears to exist a value scale on which a minimum value could be indicated under this value segmentation would not be possible, in a particular language.

m

In case of a comprehensive vocabulary the calculation of *v* is virtually impossible without a computer. However, the sequential character of the operations in a computer is a handicap if we compare it to the strong synthesizing faculty of the human mind, so that it takes a relatively long time to execute this comples program.

The project we have discussed in this contribution started with a very simple program consisting of the elaboration and use of alphabetical lists. It has been extended so as to comprise the elaboration of a more complex algorithm capable of indicating structures in language forms in a way presumed to be analogous to the procedure used by man. The standards we applied were not known beforehand, but all of them originated from the formal comparison of successive partial results. All this could only be done by a computer. The programming language we used was PL/1 because algorithms written in this language can easily be modified when one wants to take new criteria into account. So, this research meets the requirements of the definition for "computer assisted research". The segmentations we have calculated correspond to those applied instinctively by an ordinary user of the language. They are rarely wrong; the algorithm did not yield some possible segmentations. The introduction of new criteria could solve this problem, e.g. if one were to take into account the context and the frequency of word forms and also if one were to consider phonetic filters.

This method can be applied to any French vocabulary, without any modification or preparation, and probably to any Indo-European language. Our team is doing this at present. The method can be adapted for research into other elements of different linguistic levels (syllables, syntagme).

The calculation of the linguistic structure is not something new. What seems to be new is the fact that it has been shown in what way the morphological structure is completely present in the form (all other methods, including the "structuralist approach" worked with a preliminary semantic interpretation). The proof for this was supplied by this automatic approach, which shows there is a structure in the linguistic forms. These very forms are the starting point and the result is obtained by using these forms only.

On the level we worked the language code seems to contain all information which is required to learn its structure and which is necessary for its use as a means of communication. So, in principle it would be possible to learn a language by the mere communication of it.

*

* *

This contribution, which is intended to show the possibilities of the computer in linguistic research, does not give an exhaustive survey of the problem of automatic morphological segmentation. The method we have dealt with is, in fact, much more complex and more complete than the outline we have given here. A detailed description of methods and results, together with a justification and a more extensive linguistic, methodological and psychological interpretation will be published in a book entitled :
Le morphème : une expérience de recherche linguistique automatisée.

(1) *Contribution à une étude de la surabondance en français écrit;* in Linguistics 43 (1968), pp. 5-31.

Eléments distinctifs et surabondants dans le mot français parlé, which will appear in Actes du Xe Congrès des linguistes de Bucarest.