QUELQUES REFLEXIONS A PROPOS DE L'ANALYSE SYNTAXIQUE AU MOYEN DE L'ORDINATEUR

Elie NIEUWBORG, Chargé de cours associé, U.C.L.

0. Il me paraît inutile d'insister ici sur des travaux très courants tels que la confection de listes de fréquence et de dictionnaires inverses : les réalisations du Séminaire de Langue néerlandaise dans ce domaine sont citées dans la liste des travaux qui accompagne cet article.

1. Je me contenterai de soulever ici quelques problèmes théoriques de l'analyse syntaxique au moyen de l'ordinateur. Quel que soit l'instrument d'analyse que le chercheur utilise, il devra résoudre un problème de terminologie. A supposer même que cette terminologie fût univoque, ce qui sera rarement le cas, elle constituera toujours un système fini, un sous-ensemble d'une langue naturelle qui prend un caractère restrictif et par ce fait même se rapproche des langues artificielles. Décrivant une langue naturelle, qui est non-restrictive, le chercheur se voit obligé d'adapter et d'interpréter sa terminologie, de l'étendre par des termes nouveaux, de peser et de nuancer le sens qu'il entend donner aux termes qu'il utilise. C'est un fait bien connu, que des notions trop souvent considérées comme "universelles" ne couvrent pas les mêmes réalités, même dans les langues étymologiquement apparentées : dans le système syntaxique néerlandais, la voix passive n'exerce pas la même fonction que dans le système français et un tour réfléchi français n'est pas nécessairement l'équivalent de la construction réfléchie néerlandaise. Même la relation apparemment très simple de déterminant à déterminé pose des problèmes extrêmement complexes. Il est déjà difficile de décrire cette relation à l'intérieur d'une langue : *sportpubliek* où *sport* est déterminant de *publiek* n'est pas l'équivalent du même mot *publiek* déterminé par l'adjectif *sportief*. Mais une description adéquate de la relation entre l'adjectif *sportief* et le nom *publiek* ne conviendrait pas à la même relation adjectif - nom en français, le groupe *public sportif* étant équivoque et exprimant les relations des deux tours néerlandais.

La programmation de l'ordinateur qui se fait dans une langue artificielle exige une formalisation minutieuse et précise. Le nombre de symboles de cette langue étant limité, on est en droit de se demander si une description *exhaustive* est possible. Au cours de ses travaux, le chercheur devra sans aucun doute multiplier par quelqu'artifice le nombre de symboles. Ne risque-t-il pas dès lors d'obtenir finalement un calque de la langue étudiée ? D'autre part, le même danger guette également le chercheur qui refuserait la formalisation, s'il n'est pas conscient du fait que par le choix même de sa terminologie il s'engage dans la voie de la modélisation. Il nous semble toutefois probable qu'il évitera plus facilement les écueils

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

de la formalisation s'il exploite dans sa description toutes les ressources de la non-restrictivité de la langue naturelle.

Sur un plan plus général on pourrait poser la question en d'autres termes : une grammaire formalisée peut-elle satisfaire aux exigences de simplicité ? Au fur et à mesure qu'elle devient plus exhaustive, n'atteindra-t-elle pas un degré de complexité tel qu'elle devient en réalité une transcription ou une traduction de la langue naturelle étudiée ? Supposons même que les symboles revêtent un caractère que dans le langage courant nous appelons "général" et que les éléments linguistiques puissent être décrits ou "traduits" en chaînes de caractéristiques comme celles employées généralement en GGT (+ ou − défini, + ou − individualisé, + ou − inchoatif, etc.), les linguistes pourront-ils s'entendre sur une définition universelle et rigoureuse de ces termes ? La diversité des cultures ne les obligera-t-elle pas à redéfinir ces termes par des approximations successives pour chaque langue examinée ? Si ces questions ont été soulevées ici, ce n'est pas pour essayer d'esquisser une réponse, mais pour attirer l'attention sur les aléas possibles de la formalisation et spécialement de la programmation d'une syntaxe sur ordinateur.

2. L'étude syntaxique au moyen de l'ordinateur peut se faire par deux approches fondamentalement différentes : on peut soit examiner des textes pré-analysés, soit employer l'ordinateur comme instrument d'analyse.

2.1 L'examen de textes pré-analysés présente des avantages certains par rapport à un traitement exclusivement "manuel". La rapidité avec laquelle se font les opérations permet d'examiner un nombre de paramètres très élevé en un minimum de temps. L'analyse à laquelle nous avons soumis 5.000 phrases néerlandaises avait pour but d'examiner la structure linéaire de la phrase, c'est-à-dire la succession des différents groupes (1). Il était impossible de définir au départ quels facteurs devaient être retenus : la fonction du groupe ? sa longueur ? sa structure interne ? la catégorie grammaticale à laquelle appartient la base du groupe ? Ensuite, il eût été dangereux de s'engager dans un système de description bien déterminé en négligeant tous les autres : ce système étant par définition restrictif, nous n'aurions pas examiné la structure du néerlandais, mais la comptabilité du système de description avec la structure de cette langue. Un de nos résultats le prouve très clairement : la distinction "traditionnelle" entre compléments d'objet et compléments circonstanciels est inadéquate pour l'analyse du néerlandais. Si l'on hésite à examiner "manuellement" un grand nombre de paramètres, la facilité avec laquelle le traitement électronique se pratique, invite tout naturellement à superposer différents systèmes d'analyse et à multiplier le nombre de facteurs a traiter. Une fois que l'analyse a ete faite, la tache du chercheur consistera à interroger la machine sur les rapports entre les différents paramètres et surtout à interpréter la valeur des rapports décelés par l'ordinateur.

2.2 La seconde approche, celle de l'analyse automatique est beaucoup plus complexe. A l'entrée, la machine est alimentée d'une succession de symboles ou groupes de symboles séparés par des "blancs" :

(1) Cfr. bibliographie ci-dessous, 1.

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

ces blancs étant également comme des symboles. Qu'attendons-nous a la sortie ? Le résultat maximal est une analyse syntaxique complète. Mais en quels termes celle-ci sera-t-elle exprimée, si ce n'est dans les termes mêmes de la programmation ? Si cette analyse est exhaustive et exacte, nous avons la preuve que nous disposons d'une grammaire capable de traiter les données d'entrée. L'auteur de cette grammaire n'est pas l'ordinateur, mais le chercheur qui, par le truchement de la machine, a trouvé une confirmation des hypothèses grammaticales qu'il avait formulées. Alimentée des symboles tels qu'ils sont décrits ci-dessus, que pourrait découvrir la machine sans cette grammaire ? Son programme pourrait lui demander de comparer entre elles les chaînes de signes séparées par des blancs. Elle découvrira que certaines chaînes ont une très grande fréquence et que d'autres sont très rares, que la longueur de ces chaînes est variable et que les différentes longueurs ne sont pas réparties de façon égale. Si elle dispose d'un outil statistique suffisamment puissant, elle pourra calculer les distributions et éventuellement les comparer à des distributions théoriques. Elle nous dira également que certaines chaînes sont des sous-ensembles (phénomènes que nous décrivons comme flexion, dérivation, composition). Si elle analyse le français, elle découvrira peut-être que souvent plusieurs chaînes se terminant par le symbole S se succèdent et qu'à proximité de celles-ci se trouve souvent une chaîne qui se termine par ENT ou que la chaîne LA est souvent suivie d'un ensemble qui se termine par ION ou par EE. Procédant ainsi par comparaisons successives, elle finira par découvrir les paradigmes des ensembles de symboles dont on l'a alimentée. Elle pourrait même se constituer de la sorte une grammaire taxinomique opérationnelle. Mais rien ne l'empêchera de construire une phrase telle que : Le pigeon mort gagne le prix de Rome et son pain en travaillant dans une usine à Bruxelles. Dans une grammaire taxinomique basée uniquement sur des symboles vidés de leur rapport avec les données de l'expérience cette phrase n'est pas invraisemblable. Elle serait basée sur les groupes suivants : Le pigeon gagne le prix de Rome / Le pigeon mort/ L'ouvrier gagne son pain en travaillant dans une usine à Bruxelles.

La phrase que l'ordinateur aurait produite de cette façon est-elle grammaticale ? Elle l'est certainement dans le contexte de la grammaire utilisé. Mais il s'agit ici — comme nous venons de le dire — d'une grammaire de symboles vidés de leur rapport avec les données de l'expérience, ce qui la rend inutilisable aux fins de la communication humaine et inadéquate en tant que description linguistique. Admettons qu'elle puisse reconnaître de *façon infaillible* les groupes syntaxiques dans la phrase — ce qui est très peu probable —, et qu'elle ait donc repéré les symboles qui indiquent les liens syntaxiques, elle ne pourra toutefois fournir aucune information concernant la *nature* de ces liens.

Il est évident que la machine ne pourra découvrir que les liens syntaxiques marqués dans les données d'entrée. Dans les langues à flexion réduite (telles que le néerlandais, l'anglais, le français) ces marques font souvent défaut. En outre, les homographes seront des sources de nombreuses erreurs de classement : en français une forme verbale précédée du pronom sera traitée par la machine comme un nom précédé de l'article *le*. Pour la plupart des langues, la grammaire taxinomique construite par l'ordinateur sera très rudimentaire, si le programme ne contient pas de règles de levée d'ambiguités. L'élaboration de ces règles

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

demande de la part du linguiste une attention particulière pour tous les éléments des chaînes d'entrée. Mais le linguiste ne se contente pas d'une taxinomie, celle-ci fût-elle parfaite. Il désire également analyser et interpréter les liens qui unissent les données du langage. A cet effet, il "apprendra" à l'ordinateur un ensemble de règles, une grammaire.

Il est très difficile à l'heure actuelle d'évaluer la portée linguistique de ces règles. Il serait important de savoir dans quelle mesure le programme d'interprétation linguistique imposé à l'ordinateur simule le comportement humain. En principe, l'ordinateur exécute les opérations suivantes : consultation du dictionnaire composé généralement d'un mélange de formes, de lemmes et de radicaux opérationnels, analyse et interprétation des formes, levée des ambiguités. La succession des opérations est déterminée par des considérations d'économie. Ainsi il est évident que la forme verbale française *a* se trouve dans le dictionnaire, mais que *trouvait* sera reconnu seulement après analyse et que le dictionnaire contiendra le radical *trouv* et la désinence verbale *ait*. Il est fort probable que l'homme se comporte de la même façon. Mais ces dictionnaires sont des listes opérationnelles qui bien souvent n'ont plus rien de linguistique. Dans notre programme expérimental de lemmatisation des formes verbales néerlandaises, notre première opération consiste à détacher le préfixe du participe passé *ge* du radical. Cela pose évidemment des problèmes pour les verbes dont l'infinitif commence par *ge*. Ainsi, pour pouvoir analyser les formes de *geven*, nous avons dû prévoir les "radicaux" suivants : *ef* (geef), *eft* (geeft), *ven* (geven), *geven* (gegeven), *gaf* (gaf), et *gav* (gaven). Le programme d'analyse devient un amalgame de règles linguistiques et de règles opérationnelles.

Malgré des inconvénients évidents, l'analyse automatique – même si elle risque de n'être jamais exhaustive – reste une opération payante du point de vue purement linguistique. L'ignorance totale de la machine oblige le linguiste à expliciter toutes les règles d'analyse (malheureusement aussi les règles purement opérationnelles) de la façon la plus univoque, car la machine ne tolère aucune imprécision. Le chercheur doit exploiter jusqu'à la limite les éléments formels du langage et arrivé au bout de son analyse formelle, il devra étudier dans quelle mesure certains liens syntaxiques sont marqués par des éléments sémantiques, psychologiques ou socio-culturels et il introduira ceux-ci dans son programme. Il est impossible de prévoir quelle quantité d'information le linguiste devra "apprendre" à l'ordinateur. Combien d'instructions ne faudra-t-il pas donner à l'ordinateur pour l'empêcher d'interpréter le groupe "toute la soirée" comme complément d'objet dans "Il a écrit toute la soirée", alors que cette interprétation est correcte dans "Il a décrit toute la soirée "? Les instructions ont-elles encore un caractère linguistique ? Les éléments dont la machine a besoin ne seront-ils pas une paraphrase de la langue analysée, comme nous l'avons déjà suggéré plus haut ? Quoi qu'il en soit, nous croyons que le défi que la machine a lancé au linguiste sera profitable au chercheur qui n'abdique pas devant les difficultés de la programmation.

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

Publications préparées au moyen de l'ordinateur.

1. E. NIEUWBORG, *De distributie van het onderwerp en het lijdend voorwerp in het huidige geschreven nederlands in zijn A.B.-vorm,* Plantijn, Antwerpen, 1968, 524 p.

2. E. NIEUWBORG, *Retrograde Woordenboek van de Nederlandse Taal,* Plantijn, Antwerpen, 1969, 1115 p.

3. *Frequentielijst van het geschreven Nederlands.*
   A paraître dans la série "Travaux de la Faculté de Philosophie et Lettres de l'Université Catholique de Louvain. Série Microfiches".

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

SUMMARY

Computers can easily be used to classify and compare preanalysed sentences in order to examine distributional characteristics of a natural language. On the other hand, computational syntactic analysis is a very complex problem. Adequately programmed, a computer — if powerful enough — is not unlikely to discover a — mostly very trivial — taxonomic grammar of a natural language. This type of grammar however would be of little use to the linguist who is concerned with language as a system of human communication. Programming an exhaustive grammar is impossible without introducing semantic rules and might even prove to be impossible. Linguists however should take this risk.

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