GIACOMO FERRARI

DICTIONNAIRE AUTOMATIQUE ET DICTIONNAIRE-MACHINE: UNE HYPOTHÈSE

1. Pendant quatre ans le groupe de recherche linguistique du *CNUCE* a produit un dictionnaire automatique de la langue italienne, c'est-à-dire une liste de toutes les formes, lemmatisées et enregistrées sur mémoire périphérique.

Pour cette production on a employé:

a) une liste d'entrées décrites soit sous l'aspect lexical, soit sous l'aspect morphologique;

b) des programmes de flexion portant soit sur des règles générales, soit sur l'interrogation de codes morphologiques particuliers de l'entrée.

Etant donné qu'un tel dictionnaire a été projeté pour être employé dans les procédures de lemmatisation automatique, à l'heure actuelle il est utilisable, bien qu'il puisse être considéré terminé ou non terminé suivant le degré de précision de la lemmatisation que l'on veut obtenir et le degré de sophistication des procédures que l'on veut employer pour lemmatiser.

Mais on s'attend aussi à ce qu'il soit employé comme partie d'une machine pour l'analyse linguistique d'un texte. Cet emploi demande une prédisposition du dictionnaire à fournir non seulement des données sur l'accord morphologique dans la phrase, mais aussi des traits de sélection qui permettent de réduire le nombre trop élevé de séquences possibles à un nombre plus proche de celui des phrases grammaticales de façon à nous mettre en condition d'exercer un choix plus fin entre arbres alternatifs.

2. Il est évident que le projet n'est qu'un perfectionnement des deux moyens déjà employés dans la production de notre dictionnaire. D'un côté on devra généraliser les programmes de flexion en en faisant une structure capable d'analyser. Ce n'est pas notre intérêt ici de parler de cette branche du projet; il suffit de dire que la généralisation des

règles morphologiques entraîne un processus de précision aussi des codes morphologiques attachés aux entrées.

Le projet que nous sommes en train de développer est, au contraire, la recherche et l'assignation des traits sélectifs et des codes syntaxico-sémantiques des entrées ce qui n'est qu'un raffinement des codes lexicaux que nous avons employés dans la phase précédente.

Nous justifierons ensuite notre confusion entre traits lexicaux et traits syntaxiques, mais je voudrais souligner, avant de commencer, l'importance de deux autres problèmes:

a) l'économie d'espace, que nous devrons réaliser non seulement avec des moyens techniques, mais aussi en essayant de concentrer et de distribuer l'information d'une façon plus simplement utilisable;

b) l'organisation des entrées qui devra refléter nécessairement la disposition de l'information pour réaliser le principe d'économie.

3. Pour nous mettre au travail il nous faut partir d'une série d'observations et de postulats.

3.1. La première observation est que nous sommes partis d'un travail lexicographique et que nous sommes obligés de continuer à travailler sur le lexique. Nous aurions eu une autre solution: nous aurions pû essayer notre dictionnaire, mot après mot, dans de petites phrases correspondant à une syntaxe très réduite et simple, en classant chaque mot suivant son fonctionnement dans le contexte. Mais l'extension du dictionnaire rend presque infini le nombre des combinaisons. D'autre part, en travaillant avec une petite syntaxe inventée dans ce but, on a la certitude que notre classement aura les limites de la syntaxe même. En effet, un dictionnaire exhaustif ne peut être exhaustivement traité et classé si ce n'est au moyen d'une syntaxe exhaustive.

On a pensé, donc, de renverser l'ordre des choses et de continuer à décrire le vocabulaire en cherchant, parallèlement, les règles syntaxiques capables de traiter les catégories que nous trouvons.

3.2. Pour justifier cette position, nous acceptons le postulat que chaque article d'un dictionnaire quelconque ne soit que la neutralisation d'un ou plusieurs types de contextes de l'entrée.

Suivant ce postulat il sera possible de dégager une série de traits syntaxiques par la simple interprétation d'un article de notre dictionnaire de base. Notre confusion entre données lexicales et données syntaxiques se justifie par le fait que nous employons un article comme

image d'une série de contextes et la série de contextes comme image d'une structure syntaxique.

3.3. Ce principe entraîne un autre postulat. La définition que le dictionnaire attache à chaque entrée, d'un côté a la fonction de nous faire comprendre le signifié d'un mot, de l'autre représente effectivement les traits contextuels que nous cherchons. Elle appartient, donc, au langage, puisqu'elle nous renseigne sur une certaine réalité, parfois concrète, parfois linguistique, et à un métalangage, puisque chaque élément, ou certains éléments les plus significatifs, peut être considéré comme le symbole d'un certain trait.

Si, par ex.,

BEAUTÉ est défini « Qualité de ce qui est beau », nous sommes renseignés sur ce qu'est la « beauté », mais nous savons aussi que le mot *BEAUTÉ* appartient à la classe *qualité* et aura une certaine série de compatibilités définies par cette classe.

De même, si

CHIEN est réellement un « Mammifère domestique des Carnivores » il est vrai aussi que le mot *CHIEN* appartient à la classe linguistique des *mammifères* et, donc, aura, parmi la classe des *animaux*, les compatibilités caractéristiques des *mammifères*, tel que, p. ex., la famille des verbes qui comprend *ACCOUCHER*, *POULINER*, *VELER*, *METTRE BAS* etc.

4. Evidemment, sur cette route se posent déjà des difficultés. La première est que, ayant choisi un dictionnaire de base, il faut en décoder le métalangage et distinguer les différents niveaux d'analyse lexicale et syntaxique. En effet, suivant notre expérience, nous savons que les explications peuvent avoir la forme d'un synonyme, d'une description ou simplement d'une relation syntaxique par rapport à une entrée-base. La question se pose, donc, de savoir s'il s'agit de substances différentes, de différents niveaux d'analyse ou, plus simplement, de moyens différents arbitrairement employés pour décrire la même chose.

Or, si nous acceptons le point de vue optimiste qu'un dictionnaire, en tant qu'il représente une certaine tradition lexicographique, emploie des moyens assez réguliers, formalisés et, par conséquent, déchiffrables et immédiatement employables comme symbole univoque d'un métalangage unitaire, les difficultés seront résolues par un processus de décodage, interprétation et traduction en symboles mécaniquement utilisables.

*

5. Pratiquement le projet se développe en trois phases dont nous sommes en train d'accomplir la première:

a) Nous transcrivons, en correspondance de chaque entrée, la définition, ou les définitions, qui lui sont attachées par le dictionnaire, et toutes les autres indications qui peuvent avoir une fonction de sélection préalable, telles que l'indication de langage spécialisé, de procédé de déplacement sémantique etc.

La transcription de la définition doit être faite suivant certaines règles qui visent à la réduire à l'essentiel en en mettant en évidence les termes les plus significatifs. Il s'agit d'une significativité par rapport à l'article même et non pas par rapport à l'hypothèse générale. En effet, on a inventé des règles de lecture et d'interprétation formelle de l'article, justement pour empêcher toute intervention subjective. Surtout l'article doit être lu simplement comme un article de dictionnaire sans aucune conscience de la double fonction dont nous avons parlé plus haut.

D'autre part, on emploie trois types de formats de transcription, suivant les trois types de définition grossièrement décrits.

b) Successivement on se propose d'essayer des groupements grossiers de mots suivant leurs explications, ou, pour se rattacher à l'hypothèse, suivant leurs traits syntaxico-sémantiques. Le système plus général devrait consister à trier une première fois les définitions, de façon que l'on puisse observer, d'un côté, d'éventuelles déviations du métalangage fixé par le dictionnaire de base, et, de l'autre, les éléments constants de la définition.

A partir de ce moment on pourra commencer à faire un travail originel, parce que seulement ici on aura décodé le métalangage du dictionnaire et on pourra essayer d'économiser les éléments constants des définitions, en commençant à les utiliser réellement comme des traits sélectifs par rapport à un contexte.

Avant de continuer il faut souligner que, bien qu'ici on ne parle que d'une méthode générale, on est conscient qu'il sera nécessaire d'aborder le travail avec plusieurs systèmes d'analyse. En effet, l'emploi des différents formats de transcription a justement la fonction de faciliter une première division du travail suivant une certaine hypothèse que nous avons énoncé. De même, je n'ai pas parlé, jusqu'ici, de classement ou d'encodage parce que notre expérience nous rend sûrs que seulement le premier tri nous donnera les connaissances suffisantes pour projeter les procédés corrects.

Aux lexicographes je rappelle l'importance même de la première partie de ce projet qui nous libère de l'esclavage de la consultation du dictionnaire pour toute opération de lemmatisation.

c) Nous passons donc à la troisième phase qui est complètement hypothétique. Nous sommes sûrs, du moins, que nous aurons la possibilité de traduire certaines séquences en codes généraux appartenants à un métalangage plus abstrait que le métalangage que nous avons reçu du dictionnaire. Nous savons, d'autre part, que cette classification n'est pas suffisante pour la création d'un instrument d'analyse assez puissant pour être exhaustif. Pour citer un exemple, si la définition de *BEAUTÉ* est « Qualité de ce qui est beau », je peux lui attacher le code indiquant qu'il s'agit d'un abstrait, ce qui constitue déjà une sélection, mais ne me renseigne pas sur le fait qu'il s'agit d'une *qualité*, par opposition, p.ex., à *sentiment*, *vertu* etc.

L'objection la plus facile serait pourquoi ne pas attacher un code général indiquant *abstrait* et un code plus particulier qui serait la traduction immédiate du premier terme de la définition, *qualité*, *sentiment*, ou *vertu*. Ceci est sans doute vrai, mais ne résout le problème qu'à moitié. Il existe, en fait, une large quantité de catégories représentées par un ou deux mots dont la traduction en codes serait peu économique.

Mais si nous revenons un instant à notre postulat suivant lequel nous traitons un métalangage qui n'est que l'emploi métalinguistique du langage, nous voyons que nous n'avons pas besoin de tellement de codes. Il sera suffisant de renvoyer une classe de mots au mot du dictionnaire qui constitue le premier élément de la définition. Le même processus peut être appliqué aux éléments successifs, s'ils apparaissent suffisamment significatifs.

A la fin de ce travail on aura renfermé le vocabulaire dans un réseau de branchements constitué par:

a) une série de catégories générales qui représentent une première division en branches du vocabulaire;

b) des branchements exprimant des rapports hiérarchiques entre les mots du vocabulaire;

c) des branchements exprimant des rapports d'équivalence entre certains mots.

6. On arrive, donc, à concevoir le dictionnaire comme un arbre dont les premiers noeuds sont des catégories générales et les noeuds inférieurs sont les mots du vocabulaire hiérarchiquement rangés.

La caractéristique fondamentale de cet arbre est que chaque noeud peut être constitué d'un ou plusieurs mots, de même qu'un mot peut représenter un ou plusieurs noeuds.

Il reste encore un point à éclaircir, l'emploi de cet arbre, étant donné surtout qu'il s'agira de plusieurs arbres, du moins un arbre pour la partie nominale du discours et un arbre pour les verbes.

Il est évident qu'en phase de recherche syntaxique il sera assez facile de descendre les arbres et vérifier une règle sur deux ou plusieurs branches que nous savons déjà compatibles entre elles. En termes plus pratiques, il sera facile de prendre tous les noeuds équivalents des arbres produits en traitant par une règle seulement ce qui se trouve au dessous de ces noeuds.

Mais le contraire est vrai aussi: étant donné une phrase, une syntaxe et une table de compatibilité entre les sous-branches des arbres, on pourra limiter le dictionnaire pour l'analyse de la phrase aux sous-branches compatibles avec la sous-branche du premier mot, de façon que si dans la recherche mécanique un mot manque, la phrase n'aura aucun sens, ou le mot sera employé dans un signifié que nous ne connaissons pas.

7. Seulement dans ce sens nous réaliserons l'économie d'espace et d'information dont nous avons parlé. Nous aurons la possibilité de travailler sur des petites tranches de vocabulaire à la fois. D'autre part, si, une fois l'analyse terminée, on a besoin d'en classer les éléments lexicaux, on n'aura qu'à remonter les arbres en en mémorisant les noeuds, pour avoir le classement et une sorte d'explication.