Un système d'analyse lexicale du français

Gérald PURNELLE

SILBERZTEIN (Max) : 1993, *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX* (Paris : Masson, Collection « Informatique linguistique ») 233 pp. ISBN 2–225–84157–8.

En ouvrant cet ouvrage, le lecteur ne se fiera pas à son seul titre, qui, privé du sous-titre, serait trompeur. Il ne s'agit pas, en effet, d'un essai contenant des réflexions générales ou théoriques relatives aux dictionnaires électroniques ou à l'analyse automatique de textes, ni d'un examen des différentes méthodes utilisées dans ces domaines. Le système d'analyse lexicale de textes français présenté par Max Silberztein est en fait un maillon dans une chaîne, une pierre d'angle dans un édifice plus vaste, développé par le Laboratoire d'Automatique Documentaire et Linguistique (Université Paris 7).

En amont de cette réalisation, on trouve une ambitieuse entreprise de description et de formalisation de la langue française, à laquelle se sont attelés des linguistes, et dont procède le système d'analyse; en aval, d'autres systèmes à réaliser (un analyseur syntaxique) et les multiples usages auxquels est destiné l'ensemble.

Après avoir signalé les « outils informatiques » dont il a disposé (expressions rationnelles, automates finis, graphes), l'A. décrit d'abord les dictionnaires électroniques qui ont été patiemment constitués dans le Laboratoire, sur la base des travaux des linguistes, et dont l'ambition est d'être exhaustifs : un dictionnaire des mots simples et un autre des mots composés (avec codage grammatical de chaque mot); deux autres dictionnaires regroupant les formes fléchies des mots simples et composés. Il expose ensuite le fonctionnement de l'analyseur lexical qui, utilisant ces instruments, doit produire les données nécessaires à un analyseur syntaxique (non décrit dans l'ouvrage).

La spécificité du système INTEX réside, me semble-t-il, dans la place réservée au concept de mot composé. On sait quels problèmes posent les mots comme *chemin de fer* ou *pomme de terre* dans l'analyse linguistique et dans la mise au point de procédures automatiques. Le système INTEX élargit considérablement le concept de mot composé,

✉ Université de Liège; Centre Informatique de Philosophie et Lettres; Laboratoire d'Analyse Statistique des Langues Anciennes; 32, place du 20–Août; B–4000 Liège (Belgique).
Fax : + 32 41 23 25 45 et + 32 41 66 57 02 E-mail : u0013a1@vm1.ulg.ac.be

en intégrant comme tels dans ses dictionnaires des séquences telles que *aussi longtemps que*, *pied noir*, *table ronde*, mais aussi *aide sociale*, *après mûre analyse* ou *abandon de poste*. Ceci, qui peut paraître déroutant, a pour conséquence de gonfler énormément les dictionnaires et d'alourdir les tâches d'analyse linguistique et de codage grammatical. Mais le gain, au vu des résultats de la procédure d'analyse, est la capacité de l'analyseur lexical à lever un grand nombre d'ambiguïtés lexicales et grammaticales avant le recours à l'analyseur syntaxique.

Le système INTEX ainsi décrit, et plus précisément son analyseur lexical, est propre à rencontrer plusieurs usages, dont l'A. décrit les plus simples et les plus directs : la vérification orthographique automatique, l'indexation automatique et la production de concordance. Il évoque également les applications plus larges dans lesquels le système pourrait s'intégrer : la traduction automatique et l'interfaçage en langue naturelle. Dans de telles perspectives, la position accordée par le système INTEX aux mots composés et la façon dont il les conçoit indiquent que le niveau sémantique du vocabulaire est pris en compte, de manière détournée, dès l'analyse relevant du niveau lexical. La lourdeur de cette démarche, évoquée plus haut, semble être le prix à payer pour disposer d'un système capable de traduire un texte sans buter en fin de traitement sur les problèmes sémantiques.