

## INGENIERIE DES LANGUES

J-M. PIERREL (Dir.)

Hermès Science, Paris, France

2000 — ISBN 2-7462-0113-5 — 354 p.

### Compte-rendu

Il y a maintenant plus de cinq ans paraissait — sous l'égide de la *National Science Foundation* américaine et la DG XIII (direction à la recherche) de la Communauté Européenne — un ouvrage collectif (Cole *et al.*, 1995) intitulé *Survey of the State of the Art in Human Language Technology*. Rédigé par un aréopage de chercheurs internationalement reconnus, ce livre ambitionnait de dresser une synthèse des recherches menées dans le domaine en pleine expansion des technologies langagières. Disponible librement sur Internet, il connut rapidement une large diffusion et constitue toujours à l'heure actuelle un point d'entrée incontournable en la matière. Ce livre est pourtant loin d'être exempt de reproches, et d'aucuns eurent beau jeu de relever le caractère très inégal des différentes contributions qui y sont regroupées (Akman, 1999). Ces critiques semblent d'ailleurs implicitement acceptées par l'éditeur en chef de l'ouvrage, pour qui le principal intérêt du livre réside dans sa diffusion gratuite (Cole, 1999).

Avec *Ingénierie des langues*, publié sous la direction de Jean-Marie Pierrel, l'éditeur Hermès Sciences nous propose une synthèse francophone qui relève des mêmes objectifs. La diffusion de cet ouvrage n'étant bien entendu pas libre, le lecteur peut légitimement prétendre à une qualité rédactionnelle et scientifique supérieure. Rassurons-le tout de suite : cette qualité est au rendez-vous.

En particulier, il faut rendre hommage au travail éditorial effectué par Jean-Marie Pierrel. Bien souvent, les ouvrages collectifs pèchent par manque de cohérence. Le *Survey of the State of the Art in HLT* ne fait pas exception en la matière, qui arrive à présenter à la suite des thématiques aussi différentes que la traduction automatique, la multimodalité et enfin le traitement du signal ! Au contraire, l'état de l'art proposé dans *Ingénierie des langues* repose sur une structuration tripartite cohérente qui s'avère bienvenue du point de vue des développements récents du domaine.

Tout d'abord, une première partie<sup>21</sup> fait le tour des techniques génériques utilisées aux différents niveaux de traitement de la langue (lexique, syntaxe, sémantique et pragmatique). On peut regretter l'absence de références aux théories formelles — ou statistiques — sous-jacentes à l'élaboration de ces techniques<sup>22</sup>. Il n'en reste pas moins que cette partie introductive donne un aperçu complet des recherches en la matière, sans se

soucier pour l'instant de leur application sur des problématiques précises.

La seconde partie de l'ouvrage est intégralement consacrée à la création et à la gestion de ressources linguistiques. La place conséquente (100 pages) accordée à cette problématique témoigne de l'importance qu'ont pris les méthodes empiriques basées sur les données en ingénierie linguistique. On notera à ce sujet que cette évolution se retrouve en linguistique pure, où l'on voit se développer — après des années d'introspection chomskyenne — une linguistique dite "de corpus" (RFLA, 1999). Cette prééminence explique sans doute que l'ouvrage traite avant tout de la mise en œuvre de corpus textuels : annotation morpho-syntaxique ou autre, alignement de corpus multilingues, normalisation des ressources linguistiques (XML, TEI SGML). À l'opposé, l'extraction et la constitution de lexiques ne sont étudiées<sup>23</sup> à la marge que du point de vue à forte valeur ajoutée de la terminologie (constitution de thesaurus, d'index structurés, d'ontologies...). Il s'agit pourtant d'une problématique non négligeable si l'on songe par exemple à la tendance à la lexicalisation forte des formalismes d'analyse syntaxique actuels (lexique-grammaires, TAG, LFG, HPSG, grammaires de dépendances).

Enfin, la dernière partie fait le bilan des principaux domaines d'application de l'ingénierie des langues (construction de ressources terminologiques, recherche d'information textuelle, résumé automatique, traduction assistée par ordinateur, compréhension et génération automatique de textes et enfin dialogue homme-machine). S'appuyant sur les techniques génériques présentées dans les parties précédentes, chaque chapitre propose un bilan généralement bien conduit et très lucide du domaine considéré : ne sont en particulier éludés ni les problèmes rencontrés ni le caractère limité des applications industrielles réellement développées. De ce point de vue, il est frappant de remarquer que si l'ingénierie des langues donne de plus en plus lieu à des transferts de technologie fructueux, les applications qui en résultent reposent encore sur des techniques relativement grossières<sup>24</sup> ne faisant appel au mieux qu'à une analyse linguistique de surface très locale. Les succès actuels de l'ingénierie des langues ne doivent donc pas masquer la distance qui nous sépare de systèmes robustes, efficaces ou conviviaux reposant sur l'utilisation de traitements linguistiques profonds. Un des mérites de cet ouvrage est précisément de relever les apports qui peuvent être raisonnablement attendus

<sup>21</sup> Intitulée "outils et formalismes pour le traitement des langues", elle reprend des contributions révisées et réactualisées du numéro "État de l'art" de la revue T.A.L. (38(2), 1997), désormais publiée également chez Hermès.

<sup>22</sup> À la manière du chapitre 11 ("Mathematical Methods") du *Survey of the State of the Art in HLT*.

<sup>23</sup> En partie 3, chapitre 9 : Construction de ressources terminologiques, par D. Bourigault et C. Jacquemin.

<sup>24</sup> Citons par exemple les textes à trous en génération, les méthodes par extraction en résumé automatique, ou les modèles statistiques (vectoriel, booléen pondéré) en recherche d'information.

d'une meilleure modélisation linguistique, mais aussi de nous présenter pour chaque problématique des pistes prometteuses, voire déjà opérationnelles en laboratoire.

À la différence du *Survey of the State of the Art in HLT*, cet ouvrage de synthèse se caractérise également par l'homogénéité de ses contributions<sup>25</sup>. En règle générale, chaque présentation, longue d'une vingtaine de pages, suit une structure similaire. Une introduction présente tout d'abord la problématique concernée ainsi que ses enjeux économiques et scientifiques. Suit alors un rappel historique bienvenu, qui, par l'exposé des succès ou des échecs du passé, met clairement en situation les recherches actuelles. Un survol de l'état de l'art est alors donné, qui rend le plus souvent compte des applications industrielles existantes. Enfin, une conclusion détaille — le plus souvent objectivement — les limites actuelles du domaine, ses enjeux clés et présente plusieurs pistes prometteuses pour l'avenir.

Ces présentations ne sont, faute de place, que très peu détaillées d'un point de vue technique voire même linguistique. Ne sont en effet esquissées le plus souvent que les idées clefs sous-jacentes à chaque approche. Cet ouvrage propose donc un survol très complet mais aussi très général, qui sera à même de satisfaire l'industriel cherchant à se faire une première idée des recherches du domaine, mais également le chercheur désireux, par exemple, de mieux saisir les enjeux d'une problématique de recherche proche de la sienne. Enfin, il me semble qu'il constituera un ouvrage introductif de référence pour tout doctorant s'intéressant au traitement automatique de la langue. On aurait pu craindre que le caractère très général de l'ouvrage ne nuise à une bonne compréhension de la part de lecteurs encore relativement candides. L'unanimité avec laquelle mes thésards ont accueilli cet ouvrage montre clairement qu'il n'en est rien. De ce point de vue, on relèvera le caractère très complet des bibliographies (parfois utilement complétés de pointeurs WWW) fournies à chaque fin de chapitre<sup>26</sup>.

---

<sup>25</sup> Seules quelques contributions déçoivent légèrement. On pourra ainsi regretter la part très importante accordée à des considérations purement computationnelles dans l'exposé sur la traduction assistée par ordinateur. Ce domaine d'application historique adresse pourtant des questions linguistiques essentielles pour l'ingénierie des langues. Dans un autre registre, si le chapitre consacré au résumé automatique est bien mené, son caractère égocentré sur les travaux — par ailleurs intéressants — des auteurs agace quelque peu. Enfin, les deux textes concernant sémantique et compréhension automatique ne se placent en rien dans une perspective ingénierique : présentant une bibliographie majoritairement limitée aux années 70 et 80, ces chapitres traduisent implicitement l'échec de l'Intelligence Artificielle classique face à une problématique dont il faut reconnaître l'extrême difficulté. Difficulté qui, si elle est bien rendue par les auteurs, ne justifie pas la mise à l'index implicite — voir la pirouette finale du chapitre 14 — de travaux tels que ceux de la campagne MUC (*Message Understanding System*). Certes, il s'agit de recherches très ciblées et donc à portée limitée. Mais n'est-ce pas là un moyen pour sortir de l'étude stérile de cas jouets (mal) appréhendés dans toute leur complexité sémantique et pragmatique ?

<sup>26</sup> Qu'il me soit cependant permis d'adresser un reproche sur le format des références bibliographiques. Chaque ouvrage est en effet uniquement référencé par les trois premières lettres de son auteur, accompagné de l'année d'édition. Or, cette référence très courte est hautement ambiguë. Pour ne prendre

Ainsi, la publication de cet "*Ingénierie des Langues*" est un événement particulièrement bienvenu dans l'édition scientifique francophone. Nul doute que cet ouvrage rencontrera un fort écho dans le domaine de l'ingénierie des langues mais également, souhaitons-le, en linguistique et plus généralement en sciences cognitives. En effet, s'il fait le constat des succès réels des approches empiriques basées sur les données, cet ouvrage ne manque pas d'en relever les limites. Plusieurs contributions montrent ainsi, avec pertinence et sans aucun parti pris, que le temps est venu d'un retour à des considérations plus linguistiques, voire, comme le suggèrent certains auteurs, psycholinguistiques. La linguistique computationnelle étant devenue — signe de maturité — ingénierie des langues (Cunningham, 1999), cet apport ne saurait s'envisager désormais que dans une perspective empirique et applicative claire. Comme nous l'avons relevé précédemment, l'essor de la linguistique de corpus aux cours de la dernière décennie est à même de répondre à cette attente. Acceptons donc l'augure d'une collaboration pluridisciplinaire fructueuse autour de cet objet central pour les sciences cognitives qu'est le langage. Assurément une bonne nouvelle, comme l'est la parution de cet ouvrage !

## Références bibliographiques

[Akman, 1999] Akman V. (1999). Book Review: Survey of the State of the Art in Human Language Technology. *Computational Linguistics*. 25(1). 161-164.

[Cole et al., 1995] Cole R. A. et al. (1995) Survey of the State of the Art in Human Language Technology. <http://cslu.cse.ogi.edu/HLTsurvey/>

[Cole, 1999] Cole R. A. (1999). Language technology for beginners. *Computational Linguistics*. 25(4). 641-642.

[Cunningham, 1999] Cunningham H. (1999). A definition and short history of Language Engineering. *Natural Language Engineering*. Cambridge University Press : Cambridge, UK. 5(1). 1-16.

[RFLA, 99] Collectif (1999) Grands corpus : diversité des objectifs, variété des approches. *Revue Française de Linguistique Appliquée*. Vol. 1.

---

## L'auteur de la revue critique

Jean-Yves Antoine est maître de conférences en informatique à l'Université de Bretagne Sud. Après un doctorat sur la compréhension de parole préparé à l'Institut de la Communication Parlée (Grenoble), il a mené des études post-doctorales sur le même thème au CLIPS-IMAG (Grenoble) avant de rejoindre le laboratoire VALORIA de l'Université de Bretagne Sud, à Vannes. Il dirige actuellement le groupe de recherche en ingénierie linguistique de ce laboratoire, où il conduit des travaux sur la compréhension de la parole en dialogue homme-machine, l'aide linguistique aux handicapés, l'évaluation des systèmes de dialogue ainsi que des recherches en

---

qu'un exemple, à qui attribuer la référence [BLA 91] : aux travaux de Claire Blanche-Benveniste sur le français parlé, à ceux de Philippe Blache sur les liens entre prosodie ou syntaxe, ou à ceux de Blank, Abney et al. Assurément, le lecteur novice s'y perd...

linguistique de corpus. Il anime également un groupe de recherche du PRC-I3 sur la compréhension robuste de la langue. Il est enfin rédacteur en chef de la revue en sciences cognitives *In Cognito*.