Recension de "Description linguistique pour le traitement automatique du français", par Matthieu Constant et al

Eric Laporte

Adresse de l'auteur de la recension :
Éric Laporte - 5, bd Descartes - F77454 Marne-la-Vallée CEDEX 2 - France

Matthieu CONSTANT *et al.* (éd.), *Description linguistique pour le traitement automatique du français* (Cahiers du Cental, 5), Louvain-la-Neuve : Presses Universitaires de Louvain, 2008, VI + 246 p.

Ce volume de la collection des « Cahiers du Cental » aborde des bases linguistiques du traitement automatique des langues (TAL), ce qui vaut la peine d'être signalé : les quelques volumes par an qui paraissent dans le monde sur ce thème doivent se sentir bien seuls à côté des dizaines de milliers de pages publiées dans le même temps sur le TAL statistique ou probabiliste pur... Ils peuvent se consoler de leur solitude en remarquant que l'abondance n'est pas toujours un signe de créativité ni même de diversité !

Les travaux présentés dans le volume se situent sur des chemins qui partent d'analyses fondamentales portant sur une ou plusieurs langues, qui parcourent ensuite des étapes de description, formalisation, simplification, évaluation et autres élaborations, pour enfin mener à des applications informatiques opérationnelles capables de satisfaire les besoins de consommateurs. Ces chemins sont longs et semés d'embûches. Mais les auteurs de ces treize articles font le pari que celui qu'ils tracent aboutira effectivement quelque part. Ils se situent plutôt du côté de l'étude fondamentale. Ils ne se réfèrent pas à une application précise, unique, existante et évaluable. Ils prévoient dans quel type de logiciel les résultats de leurs analyses ou de leurs descriptions, le moment venu, pourront jouer leur rôle : par exemple, des logiciels d'aide à la rédaction ou à la traduction. Le recueil traite donc plus de linguistique que d'informatique. Cependant, les auteurs manifestent leur souci de produire des résultats formels, manipulables par ordinateur.

Les articles de ce volume sont rédigés en français et traitent de la langue française. Plus de la moitié des auteurs sont de jeunes chercheurs, proches de la soutenance de leur thèse de doctorat. Les articles ont été écrits en réponse à des appels publics et été sélectionnés par un comité scientifique annoncé à l'avance.

Le lecteur ne trouvera presque rien dans ce volume qui traite directement des méthodes hybrides, celles qui combinent des informations linguistiques avec des traitements statistiques, probabilistes ou connexionnistes. Seuls un ou deux articles sur l'analyse du discours y font allusion. Cependant, pour le lecteur habitué au TAL statistique pur et désireux de diversifier sa réflexion en direction des méthodes hybrides, ce recueil pourra être une source d'inspiration. En effet, un des obstacles au développement de ces méthodes est souvent une sorte d'inhibition des chercheurs devant la complexité des structures et objets que manipule le TAL symbolique : entrées lexicales, constructions syntaxiques, segments de discours... chacun accompagné d'attributs ou de propriétés. La lecture de ce volume tend à clarifier ces notions en les illustrant par des exemples.

Les articles les plus aboutis du recueil évaluent l'exploitation d'une ressource existante pour un traitement, ou décrivent la construction effective d'une ressource linguistique : lexique, base de données..., directement utilisable, et mentionnent un nombre d'entrées. Cela concerne 4 des 13 articles.

Celui d'Isabelle Carrière répertorie des reformulations d'adjectifs relationnels en contexte : *épithélial* est paraphrasé *qui est une partie de l'épithélium* ou *qui est constitué de l'épithélium*. Ce traitement a l'avantage de caractériser finement la relation sémantique entre l'adjectif relationnel et le substantif, sans produire presque aucun métalangage, évoquant ainsi la conception harrissienne de la syntaxe et de la sémantique. L'étude vise à rendre plus informative et plus moderne la description terminologique ; elle est limitée à un domaine, mais l'analyse est rigoureuse.

Benoît Sagot et Laurence Danlos présentent l'enrichissement d'un lexique par réutilisation de lexiques et de grammaires de TAL existants. On comprend l'importance de ce

thème si l'on pense à l'énorme coût de construction de lexiques et grammaires de qualité, en main-d'œuvre et en temps. Plutôt que de construire de toutes pièces des données pour un nouveau projet, il est plus rentable d'adapter des ressources existantes. Un tel travail nécessite des compétences du même niveau, mais il est plus rapide. Cet article utilise deux sources : un lexique syntaxico-sémantique et une grammaire. Le premier est le lexique-grammaire des verbes français construit au LADL (Gross 1994). L'extraction des informations issues de ce lexique source et le passage au format du lexique cible sont des opérations complexes effectuées manuellement. En effet, les deux lexiques utilisent deux modèles de la syntaxe compatibles pour l'essentiel de leur conception, mais différents dans leur forme, et le lexique source laisse implicite beaucoup plus d'informations, que les auteurs formalisent donc pour les rendre directement utilisables. Quant à la grammaire, il s'agit d'une grammaire locale au sens de (Gross 1997) : toutes les informations syntaxiques utilisées dans la grammaire sont spécifiées au sein de celle-ci, à travers la description directe de constructions. Ici encore, les auteurs passent ces informations dans le formalisme propre au lexique cible.

Trois autres articles sont un peu plus éloignés des applications : ils décrivent l'annotation d'un corpus ou l'architecture d'un futur traitement exploitant des ressources linguistiques.

Enfin, les 6 articles restants se situent plus nettement encore sur le versant fondamental. Ils présentent une analyse préalable sur une ou quelques entrées lexicales, une méthode de construction de ressources lexicales pour le TAL, ou une ressource utile à la compréhension et à la modélisation formelle d'un phénomène linguistique. Citons ainsi l'article de Christophe Benzitoun sur la catégorisation morpho-syntaxique du mot *où*. Outre les occurrences où il analyse ce mot comme proforme ou pronom, conformément à la grammaire traditionnelle, comme dans *Où dors-tu ?*, il en distingue d'autres pour lesquels il s'agit d'une particule [1], comme dans *C'est l'époque où il faisait beau*. Vis-à-vis du TAL, l'enjeu est de disposer d'un jeu d'étiquettes morpho-syntaxiques aussi cohérent et pertinent que possible. Les propriétés observables utilisées pour distinguer les deux catégories morpho-syntaxiques sont entre autres les constructions interrogatives, et une construction à l'infinitif : *un (endroit + *jour) où sortir*.

Que le lecteur sensible à l'orthographe ne pense pas, en lisant la préface de *Description linguistique pour le traitement automatique du français*, que le logiciel de correction orthographique était en panne lors de la rédaction. La préface est simplement écrite dans la nouvelle orthographe.

Marne-la-Vallée Éric LAPORTE

Gross, Maurice, 1994. « Constructing Lexicon-Grammars ». *Computational Approaches to the Lexicon*, Oxford, Oxford University Press, p. 213-263.

Gross, Maurice, 1997. « The Construction of Local Grammars ». *Finite State Language Processing*, Cambridge, Mass., The MIT Press, p. 329-352.

[1] En anglais, *complementizer*.