Notes de lecture

Rubrique préparée par Denis Maurel

Université François Rabelais Tours, LI (Laboratoire d'informatique)

Mohamed Zakaria KURDI. Traitement automatique des langues et linguistique informatique 2. Sémantique, discours et applications. *Wiley-Iste*. 2018. 323 pages. ISBN 978-1-78405-185-3.

Lu par **Eleonora MARZI**

Université de Bologna (Italie)

Au cours de la dernière décennie, le progrès de certaines technologies – notamment l'augmentation de la puissance de calcul des machines – et une conjoncture historique particulière, qui voit la communication toujours plus orientée vers le multilinguisme et les grandes masses de données, font du TAL une discipline qui soulève de nouveaux défis. L'auteur, Mohamed Zakaria Kurdi, propose un aperçu approfondi des études existantes sur le langage à travers l'informatique, domaine interdisciplinaire par excellence. Avec une approche tant empirique que pratique, et en mettant « sur un pied d'égalité les modèles linguistiques et cognitifs, les algorithmes et les applications informatiques » ce texte parvient à donner un ample aperçu qui prend en considération les travaux classiques, mais aussi les plus contemporains.

Contenu et structure

Le livre se structure en quatre chapitres, répartis de manière équilibrée entre une approche théorique et une approche concrète. Le premier chapitre « *La sphère du lexique et des connaissances* » se concentre sur le lexique et la représentation des connaissances en introduisant la sémantique lexicale, en illustrant les bases de données lexicales et les ontologies. Le traitement des éléments de sémantique lexicale est approfondi tant au niveau de l'extension du sens, qu'au niveau de la pragmatique : Mohamed Zakaria Kurdi donne un aperçu des théories fondamentales du sens lexical, pour tout processus de catégorisation, en abordant l'approche aristotélicienne, l'approche sémique et componentielle du linguiste, Louis Hjelmslev, la sémantique du prototype de Eleanor Rosch et enfin la théorie du lexique génératif de James Pustejovsky. Quant à l'énumération de bases de données lexicales nous trouvons WordNet – avec une référence au passage à EuroWordNet, la base de noms propres Prolex, la base de données lexicales Brulex et la base Lexique. Le chapitre se conclut avec une section dédiée aux représentations formelles de la connaissance et aux ontologies, où sont illustrés les réseaux sémantiques de Quillian, les graphes conceptuels proposés par John Sowa, les

schémas de Marvin Minsky et les scripts de Schank et Abelson. En ce qui concerne les ontologies, certains projets assez représentatifs à l'heure actuelle y sont illustrés. On trouve DOLCE (*Descriptive Ontology for Linguistic and Cognitive Engineering*) et SUMO (*Suggested Upper Merged Ontology*).

Le deuxième chapitre « *La sphère de la sémantique* » propose deux approches dont l'auteur nous présente les évolutions. La première porte sur la sémantique combinatoire, qui inclut à son tour une variété d'approches comme la sémantique interprétative, la grammaire des cas et enfin la théorie sens-texte. La deuxième approche proposée est celle de la sémantique formelle qui s'est aujourd'hui assez répandue en raison du fait qu'elle se trouve à l'origine de toutes les applications dédiées au traitement computationnel du sens. À propos de cette deuxième approche, dont les fondements se trouvent dans le livre *Principia Mathematica* de Bertrand Russell et Alfred North Whitehead, l'auteur fournit les définitions de base des types de logiques les plus employées : la logique propositionnelle, la logique du premier ordre et celles d'ordre supérieur, la logique modale et la logique dynamique. Le chapitre se termine avec l'illustration du lambda calcul, théorisé en 1936 par le mathématicien Alonzo Church, et d'autres types de logiques comme celle d'ordre supérieur au premier ordre.

Le troisième chapitre s'ouvre à l'analyse du discours à travers l'apport de travaux qui vont de la linguistique à la critique littéraire, ce qui souligne encore une fois la perspective fort interdisciplinaire de cette œuvre. La première section est dédiée à l'illustration des notions de base (discours, parole, phrase, récit, texte), dans la seconde section sont présentés quelques domaines applicatifs considérés comme représentatifs, sans prétendre à l'exhaustivité, comme l'auteur le souligne sans cesse. L'illustration des notions de base se fait en référence à des travaux classiques : pour définir l'« énonciation », Mohamed Zakaria Kurdi s'appuie sur la subjectivité du langage d'Émile Benveniste, pour traiter les « déictiques » il se sert de réflexions de Roman Jakobson, et pour les « acteurs de la communication », il cite Ferdinand de Saussure. Sont également traités les notions de contexte, l'intertextualité et la transtextualité, la structuration du discours et les phénomènes de cohérence. La pragmatique occupe une section à part où l'on trouve les actes du langage de John Austin et John Searle. La seconde section du troisième chapitre est dédiée aux approches computationnelles du discours qui, comme les aspects théoriques qui les soutiennent, sont très divergentes. La variété des techniques de segmentation du discours est assez riche (base de n-grammes, réseaux bayésiens, analyse sémantique latente ou réseaux neuronaux), le cadre théorique présenté pour l'analyse automatique du discours est la théorie de la structure rhétorique, RST, (*Rhetorical Structure Theory*) et le cadre sémantique de référence est la théorie de représentation du discours, DRT, (*Discourse Representation Theory*). Le chapitre se

conclut par le traitement de l'anaphore qui occupe une section à part à cause de sa complexité.

Le quatrième chapitre « *La sphère des applications* » conclut l'aperçu en unifiant les notions théoriques linguistiques traitées tout au long de l'ouvrage avec d'autres sources de connaissances afin de construire des logiciels applicatifs appelés « linguiciels ». En particulier le chapitre aborde les linguiciels sous trois aspects : leur cycle de développement, leur architecture et leur évaluation. En ce qui concerne les architectures, on peut en imaginer plusieurs formes : la sélection de l'auteur se limite aux architectures les plus pertinentes pour le TAL. On traite les architectures sérielles, les architectures centrées sur les données, les architectures orientées objet et les architectures multi-agents. Les méthodes d'évaluation présentées sont le test structural (aussi appelé *whitebox test*), et l'évaluation de type quantitatif et qualitatif. Les applications traitées sont celles ayant un rapport avec la traduction automatique, dont l'auteur donne un aperçu historique de leurs débuts (1940) jusqu'à nos jours. Il donne aussi un aperçu des techniques employées, comme l'approche directe, l'approche par transfert, l'approche dite par pivot ou interlangue, l'approche à base d'exemple TBE et l'approche statistique. Une autre application traitée est la recherche d'information (RI) où sont énumérées les différentes approches utilisées aujourd'hui. Parmi les plus intéressantes on peut citer : les approches vectorielles et les approches fondées sur les groupements (*clustering*) qui sont également signalées par les dernières études sur le *deep learning* et sur l'intelligence artificielle. Ce trait de profonde actualité se reflète aussi à la fin du chapitre où on trouve une section dédiée au traitement des grandes masses de données et à l'extraction d'information, en particulier ayant trait à l'analyse des sentiments.

Commentaires

L'ouvrage se définit comme un aperçu de la discipline du traitement automatique des langues traitant d'une approche aussi bien théorique que pratique : les principales théories de la représentation de la connaissance et l'illustration des principales applications et bases de données existantes. L'interdisciplinarité du TAL se reflète dans l'ampleur des sujets traités dans l'ouvrage : la sémantique lexicale, la représentation des connaissances, l'analyse du discours et les applications s'entrecroisent avec des références aux travaux classiques et contemporains, ce qui montre la volonté de l'auteur de dresser aussi un aperçu chronologique.

Avec un agencement propre et méthodique, chaque chapitre est organisé de la même manière : une première partie, où sont illustrées les notions de base, est suivie par une seconde dans laquelle il est question des structures computationnelles, et des méthodes pour appliquer ces théories aux outils informatiques. Le style est clair et l'organisation du contenu aide le lecteur à suivre la variété des sujets qui vont de la

linguistique à l'intelligence artificielle, de l'informatique à la logique, de la morphologie aux transcriptions en codes.

L'ouvrage est complété par une bibliographie très exhaustive et détaillée qui tient compte de l'ampleur chronologique et thématique, et une assez claire structuration du sommaire compense le manque d'amplitude de l'index. Par rapport au paratexte, on signale la présence du sommaire du premier tome *Traitement automatique des langues et linguistique informatique 1*, qui est dédié aux notions fondamentales de la matière et qui confère à l'œuvre une complétude remarquable.

Tout aperçu porte en soi un trait spécifique, en couvrant un sujet large l'auteur est obligé de fournir des explications à propos de ses choix. Il déclare, à plusieurs occasions, son intention de passer en revue les plus importantes théories sans toutefois prétendre à l'exhaustivité. De fait, un choix s'impose, puisqu'il est impossible d'aborder la totalité des expériences existantes, le critère adopté sera celui du « plus représentatif » qui s'applique à un panorama francophone. L'ouvrage possède une perspective ample qui explique des concepts, tout en fournissant des suggestions et des sources bibliographiques pour approfondir davantage la recherche. L'ampleur des sujets traités fait ce cet ouvrage un outil précieux pour s'orienter dans l'évolution rapide et riche de la discipline du TAL.

François RASTIER. Faire sens. De la cognition à la culture. *Éditions Classiques Garnier*. 2018. 261 pages. ISBN 978-2-406-07413-7.

Lu par **Guy PERRIER**

Université de Lorraine – Loria

François Rastier propose une réorientation de la linguistique comme science sociale et de la culture. S'opposant à la conception logico-grammaticale du langage, il considère que le sens des textes n'est pas le résultat d'un calcul symbolique, mais le fruit d'une pratique interprétative sous forme d'un parcours de formes d'expression liées à des formes sémantiques. Et ce parcours met en jeu une dimension culturelle qui joue un rôle fondamental.

Pour qui veut découvrir l'œuvre de François Rastier, ce livre n'est pas le meilleur point d'entrée. J'en ai fait la douloureuse expérience. Étant totalement ignorant du contenu de ses travaux, je me suis dit, un peu naïvement, que la lecture du livre serait l'occasion de me familiariser avec eux. Or, l'univers de Rastier est peuplé d'une foule de notions étrangères au sens commun, et même à l'espace conceptuel de la plupart des chercheurs en TAL ; chaque page fait appel à ces notions qui ne sont pas redéfinies et qui nécessitent d'aller voir dans les écrits précédents de Rastier pour se les approprier. En plus, le livre est avare d'exemples qui pourraient nous en donner l'intuition. Toutefois, dans ma recherche, j'ai eu la

chance de tomber sur un texte de Philippe Gréa, « *La Perception sémantique* »[1], particulièrement pédagogique. Même si le sujet du texte ne recoupe pas complètement celui du livre de Rastier, il l'éclaire beaucoup et je voudrais lui rendre hommage. Je vous demande donc par avance de bien vouloir m'excuser de cette revue de néophyte, avec tous les manques qu'elle peut comporter. Je n'ai pas souhaité non plus être exhaustif ; le livre est très riche et je me suis arrêté sur quelques aspects qui intéressent plus particulièrement un chercheur dans le domaine du TAL.

Pourquoi ce livre ? Compte tenu des développements de ces dernières années en linguistique, Rastier a éprouvé le besoin d'actualiser ses propositions pour une réorientation de la linguistique comme science sociale et de la culture.

L'approche logico-grammaticale des langues

D'emblée, il oppose son approche de la linguistique à ce qu'il appelle l'approche logico-grammaticale, à laquelle il associe comme principaux artisans Chomsky, Fodor, Pylyshyn et Pinker. Selon cette approche, telle qu'elle est vue par Rastier, il n'y a pas de sémantique autonome des langues. Les langues sont un moyen d'accéder aux représentations du monde sous une forme logique. Dans cette conception, les signes linguistiques sont réduits à des symboles qui existent indépendamment les uns des autres avec une signification propre et immuable. Cette signification est aussi totalement séparée du symbole lui-même. Dans les textes, les symboles composent leurs significations selon le principe de compositionnalité à l'aide d'une grammaire universelle. L'interprétation d'un texte se réduit donc à un calcul sur des symboles, qui restent identiques à eux-mêmes durant tout le calcul. Cette conception donne le primat de la syntaxe sur la sémantique qui apparaît comme mécaniquement dépendante de la première. Elle est liée aussi à la conception du cerveau comme ordinateur dans sa fonction de cognition.

Rastier oppose à cela la conception saussurienne du signe linguistique dont les deux faces, le signifiant et le signifié, ne peuvent pas être séparées, le signifié n'étant pas extérieur à la langue. C'est pourquoi on peut dire que Saussure a été le premier à rendre possible une véritable sémantique linguistique.

Par ailleurs, le signe n'existe pas indépendamment du texte dans lequel il est présent. À l'opposé de l'approche logico-grammaticale, le local est déterminé par le global. Le sens d'un signe peut varier indéfiniment selon les occurrences. Il peut donc varier dans le temps et avoir une histoire.

La linguistique cognitive à la croisée des chemins

Si Rastier qualifie l'approche logico-grammaticale de cognitivisme orthodoxe, c'est parce que dans cette approche, il n'y a pas d'indépendance de la sémantique linguistique par rapport à l'univers conceptuel.

[1] https://halshs.archives-ouvertes.fr/halshs-01574243/document

La linguistique cognitive, reconnue comme telle, va plus loin en postulant que le langage n'a pas de spécificité par rapport à la cognition humaine et elle récuse toute sémantique logique en rapportant les phénomènes linguistiques à des processus mentaux. Elle a contribué à remettre en cause les postulats du cognitivisme orthodoxe. En particulier, la compréhension du sens y est plutôt vue comme une perception d'images mentales.

Rastier rapproche ces propositions des grammaires de construction, qui visent aussi à répondre aux faiblesses de l'approche chomskyenne. Dans les grammaires de construction, le rapport entre expression et sens est plus complexe, puisqu'il n'est pas défini seulement par un lexique, mais aussi par des constructions. Cependant, comme la linguistique cognitive, les grammaires de construction n'ont pas réussi à se dégager complètement du paradigme logico-grammatical : elles ne distinguent pas le signifié du concept et le sens est construit de bas en haut par composition d'unités élémentaires, alors que pour Rastier, c'est le global, le texte, qui détermine le local, le mot.

L'interprétation comme perception sémantique

Rastier, dans sa conception de l'interprétation sémantique d'un texte, s'appuie sur une relecture de Saussure. Selon lui, la présentation des idées de Saussure a été souvent tronquée : sa conception du signe linguistique ne se réduit pas à la dualité signifiant-signifié, mais le signe se définit aussi dans sa relation avec le contexte. Cela interdit de concevoir l'interprétation d'une expression comme un calcul permettant de composer le sens des sous-expressions.

Selon Rastier, le sens d'un texte est le fruit d'une perception plutôt que d'un calcul de représentations. Les unités sémantiques ne s'expriment pas comme unités discrètes relativement figées, mais comme des formes qui se profilent sur des fonds. Le sens d'un texte résulte du parcours de ces formes sémantiques. Ce parcours est évolutif, les formes se dissolvant dans les fonds par diffusion ou émergeant des fonds par sommation.

De la communication à la transmission

Après avoir abordé la question du langage sous l'angle de la cognition, Rastier montre comment on retrouve ses idées quand on l'aborde sous l'angle de la communication. La communication langagière, que Rastier appelle transmission, ne se réduit pas à un processus de codage-décodage, tel que le schéma standard de la communication le décrit. Ce schéma est parfaitement symétrique. Or, le message langagier n'est pas perçu de la même façon par l'émetteur et le récepteur.

Pour Rastier, une transmission est un fait culturel. Elle met en jeu non seulement deux protagonistes, mais aussi tout un contexte culturel. La transmission est fondamentalement une transmission culturelle, vue non pas comme un processus déterministe, mais comme une réappropriation active.

En conclusion, j'espère que ces quelques commentaires sur le livre de Rastier donneront envie à certains d'aller plus loin en se plongeant eux-mêmes dans sa lecture. Nous, chercheurs en TAL, aurions intérêt à prêter davantage attention aux

idées de Rastier, dans un souci de coller davantage à la réalité complexe des langues. La linguistique de corpus, par le rôle essentiel qu'elle attribue au contexte, parce qu'elle permet de lier expression et contenu et de prendre en compte la détermination du local par le global, est un champ d'application naturel des idées de Rastier. L'apprentissage automatique statistique est un des moyens pour le TAL d'investiguer sous cet angle la linguistique de corpus.

Maintenant, faut-il jeter à la poubelle l'approche logico-grammaticale, sur laquelle sont fondés l'essentiel de la recherche en TAL et tous ses acquis ? Faut-il construire un modèle formel alternatif fondé sur les idées de Rastier, sachant que leur très grande complexité est un obstacle à leur formalisation ? Ou faut-il enfin trouver une synthèse entre deux approches apparemment inconciliables ? L'avenir permettra de dire si le TAL a su s'approprier les idées de Rastier.