Notes de lecture

Rubrique préparée par Denis Maurel

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

Frank MORAWIETZ, Two-Step Approaches to Natural Language Formalisms, *Mouton de Gruyter*, 2003, 246 pages, ISBN 978-3-11-017821-0.

Lu par **Christian RETORÉ**

LaBRI (CNRS et Université de Bordeaux) & INRIA Bordeaux Sud-Ouest

Ce livre présente clairement ce qui est aujourd'hui connu des langages de chaînes et d'arbres nécessaires à la description de la syntaxe du langage naturel, avec une contribution essentielle : l'approche à deux niveaux. Il approfondit ainsi le lien entre les deux points de vue sur la syntaxe : model theoretic syntax *et* generative-enumerative syntax *selon la dénomination proposée par Geoffrey Pullum et Barabara Scholz. Le premier point de vue, à l'intitulé difficile à traduire en français, consiste à voir l'ensemble des chaînes ou des arbres d'un langage formel comme l'ensemble des modèles finis d'une théorie logique : certaines grammaires issues des* Head-Driven Phrase Structure Grammar*, comme les grammaires de propriétés de Philippe Blache, peuvent se voir comme un ensemble de contraintes. Le second point de vue, sans doute plus habituel en tout cas pour décrire les langages de chaînes, consiste à utiliser des grammaires de réécriture. Heureusement, cette différence de point de vue n'a pas de rapport avec la description de la langue contrairement à ce qu'on entend parfois. En effet, il arrive souvent qu'un même ensemble de chaînes ou d'arbres puisse se voir d'une part comme l'ensemble des structures satisfaisant un ensemble de contraintes et d'autre part comme l'ensemble des arbres produits par une grammaire. Un exemple classique est la classe des langages réguliers (rationnels) d'arbres : un tel langage (dont les feuilles définissent un langage hors-contexte de chaînes) peut être décrit par une grammaire régulière ou par une formule de la logique monadique du second ordre dont ils sont les modèles. Néanmoins, il est généralement admis que les langages hors-contexte de chaînes (et donc les langages réguliers d'arbres) ne suffisent pas à décrire la syntaxe du langage naturel, et il n'est pas aisé d'étendre ces résultats classiques à des formalismes syntaxiques pertinents (légèrement contextuels) : c'est l'objet de ce livre.*

L'ouvrage est découpé en trois parties. La première se compose de deux chapitres introductifs ; l'un présente les motivations linguistiques et l'autre définit proprement les objets mathématiques sur lesquels porte l'ouvrage : alphabets, chaînes, arbres, algèbres de termes typés à la William Lawvere. La deuxième partie présente les méthodes utilisées, les résultats classiques sur les descriptions dans la logique monadique du second ordre et applique cela à des fragments de *Government and Binding (GB)* de Noam Chomsky, comme l'avait fait James Rogers en

développant une description effective dans la programmation logique par contraintes. La troisième partie présente les deux étapes, appelées *lifting* et *reconstruction* et les met en œuvre pour les bien connues grammaires d'arbres adjoints (TAG) d'Aravind Joshi, ainsi que pour les grammaires minimalistes (MG) d'Edward Stabler.

Logique du second ordre monadique (MSOL) du second ordre pour la description de la syntaxe du langage naturel

Cette partie reprend et développe tout ce qui est connu de la logique monadique du second ordre pour la description de la syntaxe du langage naturel. Le premier chapitre définit la logique monadique du second ordre, en particulier sur les arbres finis avec la logique $L^2_{K,P}$ (inter-traductible avec SnS) introduite par James Rogers pour étudier des arbres des approches de type GB. Les relations de base sont la dominance et la précédence. Dans ce langage, il définit l'effet des règles, les projections maximales et quelques principes de GB. Suit un chapitre sur les machines à états finis, automates et transducteurs, tout d'abord pour les chaînes, puis pour les arbres : les automates d'arbres et les *walking tree automata*, les *top-down tree transducers* et les *macro-tree transducers*. Le chapitre suivant est consacré à la décidabilité de MSOL et à la définissabilité dans MSOL. On y établit assez joliment la correspondance célèbre entre automates d'arbres et langages d'arbres définissable dans MSOL et on présente les transductions définissables en MSOL. Le dernier chapitre de cette partie en consacré aux aspects pratiques et notamment à la mise en œuvre de GB. A priori, la co-indexation à l'œuvre dans les mouvements de GB nécessite une infinité d'indices et empêche une formulation complète en MSOL. Néanmoins, avec un nombre borné d'indices, on obtient une description MSOL d'une approximation de la grammaire qu'on peut convertir en un automate d'arbres : cette conversion et l'automate pouvant être d'une complexité démesurée, l'auteur utilise MONA et des *Binary Decision Diagrams* pour rester efficace. Des mesures convaincantes sont données, ainsi que d'autres applications de ces mêmes méthodes à la vérification de logiciels ou à l'interrogation de bases de données.

Extension de l'utilisation de la logique monadique du second ordre aux formalismes syntaxiques contextuels

La troisième partie développe un point de vue résolument original développé à Tübingen par l'auteur avec Uwe Mönnich, Jens Michaelis et quelques autres : si on ne peut décrire les arbres d'un langage linguistiquement pertinent par une formule MSOL, peut-être sont-ils l'image par une transduction définissable en MSOL d'un langage régulier d'arbres (évidemment définissable en MSOL) ? C'est effectivement le cas pour les grammaires d'arbres adjoints (TAG) — dont Mönnich a montré qu'ils correspondent aux grammaires hors-contexte d'arbres (CFTG) qui sont linéaires et monadiques — et pour les grammaires minimalistes d'Edward Stabler, en utilisant leur équivalence (sur les chaînes) avec les *Multiple Context Free Grammars* (MCFG), c'est-à-dire les *Range Concatenation Grammars* (RCG) de Pierre Boullier qui sont positives et simples.

La première étape consiste à construire l'arbre de dérivation, ce qui se fait en logique monadique du second ordre, et à calculer son *lift*, ce qui consiste *grosso modo* à le typer et à décomposer la suite des opérations à appliquer pour arriver aux arbres dérivés. Ensuite l'auteur montre comment divers types de transductions, *walking tree automata*, *macro tree transducer* ou transduction MSOL, permettent de retrouver les arbres dérivés habituels. On utilise la décomposition des opérations pour calculer les relations de dominance et de précédence entre les feuilles du *lift* qui correspondent à des feuilles de l'arbre dérivé, c'est-à-dire à des mots. Calculer dominance et précédence suffit bien sûr à définir complètement l'arbre dérivé. Concernant les formalismes linguistiques, l'auteur effectue cette transformation en deux étapes pour les TAG (vus comme des CFTG monadiques linéaires) et pour les grammaires minimalistes (vues comme des MCFGS). Il montre ainsi que MSOL, ou des automates et transductions, peuvent rendre compte de formalismes syntaxiques utilisés pour la syntaxe du langage naturel, comme les TAG ou les MG, malgré leur capacité générative non contextuelle — rappelons au passage que ces formalismes admettent néanmoins des algorithmes d'analyse polynomiaux.

Le livre conclut par les questions mathématiques découvertes en chemin et par les nombreuses perspectives que ce travail ouvre au développement de grammaires, d'analyseurs syntaxiques et de description linguistique.

Notre avis

Alors qu'on peut lire ou entendre nombre de propos obscurs sur *model theoretic syntax,* ce livre démontre la force et la netteté de ces techniques, surtout pour les langages formels qui admettent les deux types de descriptions (TAG, MG). Il souligne aussi l'importance des arbres plus que des chaînes dans la structure syntaxique des énoncés. Il prouve ainsi qu'en utilisant deux niveaux de description, on peut sortir des langages hors-contexte tout en gardant des descriptions logiques. Et, en lisant cet ouvrage, on prend conscience que bien peu est connu sur les langages d'arbres nécessaires à décrire la syntaxe des langues humaines.

On peut regretter deux manques, pourtant inévitables, concernant les grammaires minimalistes. Le premier est l'absence, dans l'ouvrage, du détail de la traduction des MG en MCFG, mais il s'agit d'un résultat important d'un autre chercheur, Jens Michaelis. On se demande aussi si cette construction peut s'effectuer avec des tuples d'arbres plutôt que de chaînes, afin de mieux connaître la structure des arbres d'analyse. La réponse est oui, mais ce résultat n'a été établi qu'en 2007 par Uwe Mönnich ainsi que par Greg Kobele, Sylvain Salvati et nous.

Ce livre est clairement rédigé, avec juste ce qu'il faut de détails, que ce soit pour comprendre la grande avancée de l'auteur, l'approche à deux étapes, ou pour profiter de sa présentation des grammaires d'arbres et de leur description en logique monadique du second ordre, notions dont il justifie pleinement la pertinence linguistique.