Note de lecture

Rubrique préparée par Denis Maurel

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

Max SILBERZTEIN. La formalisation des langues : l'approche de NooJ. *ISTE Éditions*. 2015. 425 pages. ISBN 978-1-78405-053-5.

Lu par **Noémie COLIN**

WebInterpret, France – Free-lance

L'ouvrage traite de la tâche de formalisation des langues en utilisant l'outil NooJ. Le projet de l'auteur est de décrire avec une exhaustivité et une précision absolues l'ensemble des phrases d'une langue susceptibles d'apparaître dans les textes écrits. Cet ouvrage, qui propose une description complète de formalisation des langues naturelles via *NooJ, est divisé de la façon suivante : tout d'abord l'auteur précise le projet en question et son rapport avec d'autres projets qui touchent aux mêmes domaines de linguistique formelle, d'informatique linguistique et de TAL. Ensuite, le découpage se fait en trois parties : la première qui permet de définir l'ensemble des unités linguistiques de base, la deuxième qui introduit les notions de langage formel, de grammaire générative et de machine, et enfin la troisième partie qui correspond à l'analyse linguistique automatique de textes.*

Max Silberztein présente à travers cet ouvrage le projet complet de description formalisée des langues naturelles en utilisant NooJ. Ce projet consiste à décrire avec exhaustivité et précision l'ensemble des phrases d'une langue qui peuvent apparaître dans les textes écrits. Ce projet touche différents domaines : la linguistique descriptive, la linguistique formelle, le TAL, l'informatique formelle et l'algorithme de textes. La première partie de l'œuvre constitue une sorte d'introduction et présente le cadre du projet. Il consiste à définir auprès du lecteur et du futur utilisateur de NooJ les différentes terminologies et outils de ce projet qui seront utilisés par la suite. Cela consiste également à expliquer au lecteur la distinction entre les différents domaines et applications de ce projet. L'ouvrage se présente donc sous la forme suivante : la première partie sert, comme évoqué précédemment, d'introduction au projet de formalisation ; le reste de l'œuvre est décomposée en trois parties, qui permettront de comprendre et de présenter la formalisation des langues naturelles au sein de NooJ. La première section, comprenant les chapitres 3 à 6, permet de définir l'ensemble des unités linguistiques de base, à travers la formalisation de l'alphabet, du vocabulaire et ensuite en présentant les différents dictionnaires existants à ce jour et pouvant être utilisés pour ce projet. La deuxième partie définit les notions de langage formel, de grammaire générative et de machine : tout d'abord dans le chapitre 7, sera présentée la hiérarchie de Chomsky-Schützenberger, puis les chapitres 8 à 11 expliqueront les quatre types de langages, grammaires et machines à utiliser lors de la formalisation des langues. Enfin la

troisième partie décrit l'analyse linguistique automatique de texte avec NooJ. Du chapitre 12 à 13, seront détaillées les structures d'annotation de textes, puis du 14 au 16, les trois niveaux d'analyses syntaxiques (locale, structurelle et transformationnelle) seront également détaillés et appliqués à certains exemples avec l'utilisation de NooJ. Chaque chapitre a la même structure, c'est-à-dire qu'il est composé d'une introduction suivi d'un développement, puis on trouve une conclusion pour récapituler les idées principales évoquées, suivie de quelques exercices pratiques (en utilisant NooJ ou pas en fonction des questions et de l'avancement dans la présentation de NooJ), avec des liens permettant d'approfondir certains principes, et éléments évoqués de façon non exhaustive précédemment.

Dans la première partie, l'auteur pose les bases du projet de formalisation des langues en expliquant au lecteur qu'elles résident dans les unités linguistiques. Tout d'abord, il précise que la formalisation d'une langue écrite se fait par la gestion de son système d'écriture. Après avoir évoqué tous les systèmes de codage existants aujourd'hui, il explique l'importance et la nécessité d'utiliser Unicode, même si cette solution comporte certains manquements. En effet, il faut garder en tête qu'un analyseur linguistique doit faire face à certaines spécificités de la langue, qui résident dans des variantes orthographiques, par exemple. L'auteur présente certaines solutions intéressantes qu'il faudra, bien entendu, approfondir selon les besoins du projet dans lequel le lecteur se trouve. Selon Max Silberztein, la formalisation implique de fixer le vocabulaire de façon exhaustive. Il est important de se limiter à une description synchronique du vocabulaire standard en y précisant ces éléments, c'est-à-dire les unités linguistiques atomiques, nommés « ALU » tout au long du livre. Pour reconnaître ces éléments, il faut suivre certains critères qui sont les trois suivants :

- le sens d'une ALU n'est pas complètement analysable ;
- l'usage d'une ALU est formulaïque ;
- une ALU constitue une exception à une ou plusieurs règles générales d'analyse.

Ce sont, d'après l'auteur, des critères que l'on peut considérer comme arbitraires mais qui permettent d'avoir un projet réalisable, car ils peuvent se reproduire. Les ALU sont divisibles en quatre catégories : les morphèmes, les mots simples, les mots composés et les expressions. L'auteur explique bien la différence entre ces catégories d'ALU, et énumère à travers des exemples notamment les règles qui permettent de les différencier les uns des autres. Une fois que ces éléments de formalisation sont fixés, Max Silberztein nous explique quels types de dictionnaires peuvent être utilisés dans le cadre de ce projet de formalisation. Les dictionnaires éditoriaux sont très utilisés en règle générale, mais ne sont pas adaptés à ce type de formalisation étant donné qu'ils ne comprennent pas tous les détails nécessaires à la formalisation. Les dictionnaires électroniques sont généralement bons et ont une bonne couverture, mais, encore une fois, ils ne répondent pas aux besoins de la formalisation car ils ne comportent pas une description systématique au niveau lexical, morphologique, syntaxique et sémantique. De plus, les entrées de ces dictionnaires ne sont pas toujours des ALU. Les seuls dictionnaires qui

correspondent aux besoins de ce projet sont les dictionnaires DEM et LVF de Jean Dubois et Françoise Dubois-Charlier[1].

Grâce à ces précisions sur les bases de la formalisation, l'auteur approfondit le projet en s'intéressant ensuite, dans cette deuxième partie, aux langages, grammaires et machines. Dans le chapitre 7, il introduit les notions de langage formel, grammaire générative et machine. La hiérarchie de Chomsky-Schützenberger, comporte quatre types de langages. NooJ permet de les faire interagir, ce qui n'est pas le cas dans les autres formalismes évoqués par l'auteur, tels que LEX et HPSG. L'auteur détaille ensuite chacun de ces types, dans l'ordre suivant :

1- rationnel : sous NooJ, formalisé à l'aide d'expressions rationnelles et de graphes à états finis ;
2- algébrique : utilisation de grammaires hors contexte et de graphes récursifs ;
3- contextuel : utilisation d'un système de contraintes ajouté aux graphes récursifs ;
4- non restreint : à l'aide de graphes qui permettent, comme la définition de ce type le présente, d'avoir des grammaires de réécriture les plus générales possible.

Comme expliqué précédemment, l'avantage d'utiliser NooJ pour la formalisation des langues réside notamment dans le fait que cet outil permet l'utilisation de ces quatre types de façon complémentaire.

Enfin, la dernière partie de cet ouvrage est consacré aux analyses linguistiques automatiques des langues naturelles. En effet, après avoir présenté les différents éléments de formalisation et construit la grammaire qui y est rattachée, on peut finalement construire l'analyseur et l'appliquer automatiquement à n'importe quel texte écrit. Dans cette dernière partie, l'auteur va présenter les différentes utilisations des analyses. La première, détaillée dans le chapitre 12, est la structure d'annotation de textes TAS. Ici l'auteur nous montre la puissance des phénomènes linguistiques à travers la cohabitation des différentes analyses linguistiques possibles. Il examine ensuite tour à tour ces différentes analyses. Dans le chapitre 13, il décompose l'analyse lexicale, qui consiste en une analyse orthographique, une segmentation, une analyse morphologique, et finalement l'utilisation de dictionnaires. Dans le chapitre 14, il explique l'analyse syntaxique locale, c'est-à-dire l'identification des séquences simples d'ALU. L'analyse syntaxique structurelle, qui consiste à représenter la structure de chaque phrase, et qui est importante dans certains cas pour désambiguïser dans la TAS, se trouve au chapitre 15. Finalement, au chapitre 16, l'auteur s'intéresse à l'analyse transformationnelle qui, pour lui, correspond à l'analyse sémantique. Tout au long de ces chapitres, l'auteur montre des utilisations de NooJ et des applications au TAL telles que la traduction automatique.

Pour finir, l'auteur rappelle dans une conclusion tous les points fondamentaux pour la formalisation des langues naturelles, spécifiques et nécessaires à ce projet de recherche. Il insiste également sur les avantages de l'utilisation de l'ordinateur quels

1 http://www.modyco.fr/fr/Ressources/ldlvf.html

que soient l'application ou le domaine concernés, et plus particulièrement l'utilisation de NooJ. Celui-ci permet d'utiliser des outils formels pour représenter les unités linguistiques et les grammaires qui permettent de formaliser les langues. Il est important de noter que NooJ a été développé pour cette utilisation et qu'il est également en constante évolution pour répondre aux besoins des différentes applications de formalisation aussi bien pour les linguistes que pour les informaticiens. Cet ouvrage s'adresse à toutes les personnes linguistes ou informaticiennes qui s'intéressent à la formalisation des textes écrits et/ou qui souhaitent utiliser NooJ.