Notes de lecture

Rubrique préparée par Denis Maurel

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

Bing LIU. *Sentiment Analysis and Opinion Mining*. *Morgan & Claypool publishers*. 2012. 165 pages. ISBN 9781608458844.

Lu par **Valérie BEAUDOUIN**

Telecom ParisTech

Ce livre présente une synthèse des travaux dans le domaine de la fouille d'opinion, champ de recherche apparu au début des années 2000 et qui a pour domaine principal d'application l'analyse des avis de consommateurs sur le Web. Après avoir proposé une modélisation des opinions comme un sentiment (positif, négatif ou neutre) porté par un individu sur une caractéristique (« aspect » ou « feature *») d'une entité (produit ou service) à une date donnée, le livre explore les grandes catégories d'approches : celles qui consistent à classer des documents, à classer des phrases ou à classer des opinions liées à des caractéristiques de l'entité. Deux grandes familles d'approches sont utilisées : des méthodes en apprentissage supervisé qui nécessitent au préalable d'avoir des corpus étiquetés (avis positif ou négatif ou note) et des approches non supervisées qui exploitent des lexiques de sentiments et des structures syntaxiques spécifiques. De nombreux défis restent ouverts dans ce domaine.*

Ce livre est un état de l'art sur un domaine de recherche apparu dans les années 2000, qui connaît une croissance spectaculaire en raison des enjeux commerciaux, celui de l'analyse des sentiments (*sentiment analysis*) et de la fouille d'opinion (*opinion mining*), expressions considérées comme équivalentes. Il actualise l'état de l'art de Pang et Lee, publié en 2008[1]. Ce secteur a fortement bénéficié du développement du Web 2 et de la logique participative qui conduit des milliers d'internautes à exprimer leurs opinions par écrit sur des biens, des services, des personnes, des idées. Il s'agit d'analyser des opinions exprimées spontanément en langage naturel dans des cadres interactionnels et communicationnels variés : forums, blogs, commentaires (*reviews*) sur sites, autrement dit d'extraire une évaluation sur un bien ou caractéristiques d'un bien. En ce sens, dans ces travaux, « opinion » et « sentiment » sont employés dans un sens très restreint. Ces termes entretiennent des relations de proximité et de distance avec deux autres notions : la subjectivité et les émotions. Une opinion est subjective, mais des énoncés subjectifs

1. Pang, Bo & Lee, Lillian. "Opinion Mining and Sentiment Analysis" *Foundations and Trends® in Information Retrieval* 2.1-2 (2008): 1-135.

peuvent ne pas comporter d'opinion (« je pense qu'il est rentré chez lui ») et des énoncés objectifs être porteurs d'opinion (« les écouteurs se sont cassés en deux jours »). Pour l'auteur, on glisserait de l'évaluation rationnelle à l'évaluation émotionnelle en notant très haut ou très bas.

L'auteur propose un cadre pour la modélisation des opinions qui permet de passer d'un texte libre à de l'information structurée. Dans ce modèle, une opinion est un quintuplet [cf. http://fr.wikipedia.org/wiki/N-uplet] qui comprend : une entité (le produit, le service, la personne sur laquelle une opinion est émise), une caractéristique (fonctionnalité, composante ou aspect du produit), une qualification de cette caractéristique (évaluation positive ou négative), un émetteur de l'opinion et une date. Le processus idéal consisterait à transformer un texte libre en une série de quintuplets pour ensuite pouvoir quantifier et comparer les opinions.

L'auteur distingue trois niveaux d'analyse : la classification au niveau des documents, des phrases et des caractéristiques des entités (*aspect-based sentiment analysis*).

Le classement des documents selon les opinions est le champ de recherche où les travaux ont été jusqu'à présent les plus nombreux. Les analyses portent principalement sur des corpus de commentaires ou d'avis en ligne (*reviews*). On se trouve dans un cas simplifié par rapport au modèle dans la mesure où les avis ne portent que sur une entité (produit, service) et ne sont produits que par un individu. L'objectif est d'identifier l'opinion générale sur une entité. Ce sont le plus souvent des méthodes de classification supervisée qui sont mobilisées. Les corpus d'apprentissage sont abondants puisque les avis ou commentaires sur le Web sont le plus souvent assortis de notes (nombres ou étoiles). Différents algorithmes (bayésien naïf, SVM…) sont testés et comparés, ainsi que différents traits linguistiques : les n-grammes, les mots, les parties du discours, les mots appartenant au lexique de l'opinion. Quelques travaux mobilisent des méthodes non supervisées : dans ce cas des syntagmes, potentiellement porteurs d'opinion, sont extraits grâce à des patrons morphosyntaxiques. La polarité (« sentiment orientation » ou SO) d'un document est évaluée en mesurant la distance de ces syntagmes à des mots de polarité positive ou négative comme *excellent* et *poor*, sur des corpus du Web, par exemple.

La classification des documents n'est pas toujours pertinente et il est parfois nécessaire de descendre au niveau de la phrase. Dans ce cas, les traitements se font en deux temps : d'abord identifier si la phrase est porteuse ou non d'une opinion (*ie.* phrase subjective), ensuite évaluer si l'opinion est positive ou négative. L'identification du caractère subjectif d'une phrase se fait soit en apprentissage supervisé, ce qui nécessite d'avoir au préalable un corpus annoté suffisamment important, soit par des méthodes non supervisées en mobilisant des dictionnaires de mots et de syntagmes porteurs d'opinion. La difficulté majeure de cette approche provient du fait que des phrases « objectives » peuvent être porteuses d'opinion. Pour l'identification de la polarité d'une phrase, les approches vues précédemment

sont utilisées. Une des difficultés relevées est celle du traitement des phrases ironiques.

Il est parfois nécessaire de descendre plus finement encore qu'au niveau de la phrase, parce qu'une phrase peut contenir différentes opinions sur différentes caractéristiques de l'entité (produit ou service). La fouille des opinions fondée sur les caractéristiques (*aspect-based sentiment analysis*) nécessite un recours au TAL. La première étape consiste à identifier l'entité et les caractéristiques de l'entité dont il est question, ce qui est fait à l'aide de différentes méthodes : extraction des noms ou groupes nominaux les plus fréquents, recherche de noms dans l'environnement de mots appartenant au vocabulaire de l'opinion, avec de l'apprentissage supervisé, en utilisant des *topic models*. Les points difficiles à résoudre sont liés au fait que la caractéristique de l'entité voire l'entité peuvent être implicites. La seconde étape vise à évaluer l'opinion sur chacune des caractéristiques soit en apprentissage supervisé, soit à l'aide de lexiques.

Une section est consacrée à la construction de lexiques de sentiments ou d'opinions, puisque de nombreux travaux les utilisent. Deux approches sont présentées, l'une fondée sur les dictionnaires, l'autre sur les corpus. Dans la première approche, l'idée est de partir d'un noyau de mots exprimant des opinions (positives et négatives) et d'explorer la structure synonymique et antonymique d'un dictionnaire de type WordNet pour enrichir le lexique. L'approche fondée sur les corpus vise à enrichir les lexiques en s'appuyant sur les ressources de la coordination et une hypothèse de cohérence interne des phrases en corpus. Ainsi, si deux mots sont reliés par ET, on peut faire l'hypothèse qu'ils ont la même polarité.

Ensuite quelques sous-champs de recherche sur l'analyse des opinions sont explorés : le résumé ou la synthèse d'opinions, qui peut prendre l'allure d'un tableau statistique quand les opinions sont modélisées en quintuplets ou bien emprunter aux méthodes traditionnelles du résumé automatique ; la recherche d'opinions sur le Web qui oblige à adapter les algorithmes des moteurs de recherche pour présenter les résultats selon l'opinion. Une section est consacrée à l'identification des faux commentaires qui constituent un problème de fond sur le Web, et que l'on ne peut résoudre en TAL car les faux savent parfaitement imiter les vrais commentaires. La recherche des comportements atypiques semble la plus encourageante. Enfin, un dernier chapitre est consacré à l'évaluation de la qualité des commentaires (usage en pleine expansion) en apprenant à partir d'avis de lecteurs.

Ce manuel dessine de manière très claire le paysage des travaux actuels sur la fouille d'opinion, en montrant bien les lieux où la recherche a bien avancé et ceux où beaucoup reste à faire (diversification des corpus, détection du spam…). Il est fort utile pour ceux qui s'engagent dans des travaux relevant de ce domaine. On regrettera que les travaux cités soient présentés très rapidement, plutôt sous la forme de liste de références avec peu de précisions sur les corpus, les résultats et la qualité de ces derniers.

Bernard POTTIER. Images et modèles en sémantique. *Honoré Champion*. 2012. 186 pages. ISBN 978-2-7453-2350-7.

Lu par **Dominique LEGALLOIS**

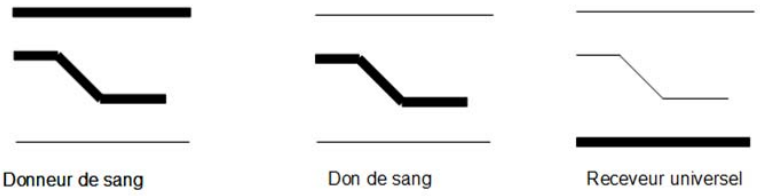
Université de Caen Basse-Normandie – Laboratoire CRISCO

Ce dernier ouvrage de Bernard Pottier vient compléter et actualiser « Représentations mentales et catégorisations linguistiques *» paru en 2000. Bernard Pottier y développe une réflexion sur les modes de représentation des faits de langue, dans une perspective cognitive originale. Le livre accorde une grande place aux représentations visuelles qui permettent d'éviter le recours à la langue usuelle.*

Commençons par quelques lieux communs, qu'il n'est pas inutile cependant de rappeler : développant la pensée de G. Guillaume, mais sachant également proposer une conception originale, Bernard Pottier a, depuis ses premiers travaux à la fin des années 40, proposé une linguistique à contre-courant. Cette linguistique peut être perçue comme un structuralisme revisité par une appréhension cognitiviste des phénomènes langagiers. Ainsi, bien que ses études en sémantique lexicale, dans les années 60, aient été érigées dans de nombreux manuels comme modèles d'analyse de la structure du lexique, Bernard Pottier s'est avant tout démarqué de la linguistique de l'époque par une approche qui a devancé, par bien des aspects, les travaux américains de la côte Ouest des années 70-80 : la linguistique doit pouvoir rendre compte de la langue comme phénomène mental, comme système sémiotique informé par notre perception et conceptualisation de l'expérience. Cette perspective motivera la première critique d'envergure du générationalisme en 1968. La recherche de Bernard Pottier, ensuite, creusera le même sillon avec des ouvrages dont les ambitions ont toujours été d'élaborer une systématicité des faits de langues, de démontrer des processus et des fonctionnements valables pour des phénomènes linguistiques différents, pour des langues différentes et parfois typologiquement éloignées. Ce même souci de systématisation est présent – et plus que jamais – dans ce dernier livre. Il s'agit, selon les dires mêmes de l'auteur, de compléter et d'actualiser l'ouvrage *Représentations mentales et catégorisations linguistiques*, publié il y a douze ans. Les lecteurs de Bernard Pottier pourraient se demander, après la somme que constituait *Représentations mentales*, quels peuvent bien être les apports novateurs de ce dernier ouvrage ; je dirais de façon un peu provocatrice, que *Images et modèles en sémantique* – petit livre de 165 pages impossible à résumer – apporte assez peu de contenu novateur pour qui connaît les travaux de Bernard Pottier. Mais paradoxalement, l'intérêt fondamental du livre réside justement dans ce retour sur les analyses sans cesse remises sur le métier, comme si l'auteur recherchait inlassablement une épure permettant de saisir la complexité des phénomènes traités. *Épure* est un mot qui convient parfaitement – un orthonyme, dirait Pottier – tant la préoccupation de la représentation la plus schématique qui soit, constitue l'objet fondamental du livre. On peut donner l'exemple de la

représentation de concepts élémentaires de sens – les noèmes – et parmi ceux-ci, plus particulièrement celui du noème de /proximité/ qu'actualisent différentes expressions : *frôler le ridicule, caresser une idée, raser les murs, friser la chute, effleurer le problème, passer à fleur de corde, tutoyer les sommets...* ce noème se représente ainsi :

Face à de telles représentations, le lecteur non averti risque fort d'être déstabilisé. Pourtant, à bien considérer les expressions étudiées, comment prétendre mieux rendre compte que par cette imagerie, la signification fondamentale qui se joue ? De même, le visuème du mot *pont*, abstrait des référents particuliers, est pertinent pour l'ensemble des emplois du mot – orthonymiques ou métaphoriques. On retrouve là, une conception partagée par les sémantiques constructivistes (théorie des opérations énonciatives, théorie des formes sémantiques, théorie des stéréotypes), même si celles-ci privilégient d'autres méthodes de représentation. Bernard Pottier va ainsi à la rencontre d'autres approches ; ne reconnaît-on pas une analyse à la façon d'*Anna Wierzbicka* (*natural semantic metalanguage*) dans le traitement de *échec :* /A vouloir P, A faire E pour avoir P, A ne pas pouvoir avoir P/ ? De même, ne reconnaît-on pas la notion de profilage de Ronald Langacker dans



la représentation suivante ?

Ces approches, en fait, Pottier les avait anticipées depuis longtemps : les primitifs conceptuels, le rapport entre latence, saillance et prégnance (inspiré de Thom) ont, très tôt, trouvé une place importante chez lui.

Le dernier chapitre du livre, intitulé *Les modèles sémantiques*, constitue une tentative révélatrice de sa démarche : proposer un modèle constituant la base des organisations cognitivo-sémantiques. Ce modèle (modèle « D ») est dit chronodéictique ; il constitue une représentation dont trois exploitations sont possibles : une exploitation excentrique polydirectionnelle fondée sur le rapport proximité-distance : [je-tu-il – les autres], [ici – là – là-bas – ailleurs], [certain – douteux – improbable – impossible] ; une exploitation excentrique axiale : [derrière – (sur, à, dans) – devant], [avant – en ce moment – après], [au-dessus – sur – au-

dessous] ; une exploitation excentrique asymptotique : [haine – indifférence – amour], [glacé – tiède – brûlant]. La représentation de ce modèle permet donc de comprendre l'organisation de champs conceptuels très divers, mais dont l'unité apparaît si l'on est attentif au jeu des profondeurs, des perspectives du schéma égocentré initial.

Images et modèles en sémantique offre ainsi une logique schématique féconde, dont chaque élément mériterait d'ailleurs d'être discuté en fonction, non seulement de sa visée analytique, mais aussi de ses propriétés pédagogiques. Ce petit livre – iconoclaste – a donc sa place sur les tables de chevet les plus différentes.

Richard MOOT, Christian RETORÉ. The logic of categorial grammars: a deductive account of natural language syntax and semantics. LNCS 6850. *Springer*. 2012. 302 pages. ISBN 978-3-642-31554-1.

Lu par **Béatrice GODART-WENDLING**

Université Paris-Diderot, UMR 7597, Laboratoire d'Histoire des Théories Linguistiques

---

La visée de cet ouvrage est de présenter – en adoptant un point de vue diachronique allant de la première syntaxe catégorielle d'Ajdukiewicz aux modèles plus récents issus du calcul de Lambek – les preuves des principaux résultats logiques qui marquèrent l'évolution des grammaires catégorielles. L'étude des propriétés formelles de ces grammaires – appréhendées comme des systèmes déductifs – conduit ainsi à retracer les écueils qui deviendront des enjeux théoriques majeurs (comme la nécessité de dépasser la limitation expressive des grammaires catégorielles résultant de leur équivalence « faible » avec les grammaires hors contexte) et à souligner l'intérêt d'exploiter les relations pouvant être établies entre certaines approches formelles (telles que la sémantique de Montague, les modèles de Kripke, la logique linéaire, etc.) et les grammaires catégorielles. Parallèlement à cet éclaircissement systématique du fondement logique caractéristique de ce type de grammaire, des analyses de quelques faits linguistiques sont entreprises et la possibilité d'user des formalismes catégoriels pour élaborer un algorithme d'apprentissage ou un parseur interactif est également mise en évidence.

Destiné aux étudiants ayant au moins un niveau de troisième année en informatique, linguistique formelle ou logique mathématique ainsi qu'aux collègues intéressés par les grammaires catégorielles et leurs fondements logiques, ce livre propose – après avoir explicité dans ses deux premiers chapitres le « socle » commun à ce type de grammaire – différents parcours permettant au lecteur d'explorer, en fonction de ses pôles d'intérêt et de ses compétences, des thématiques très diverses (touchant notamment à l'interface syntaxe-sémantique, aux choix des règles et des divers systèmes formels de représentation) en les appréhendant par le biais de la démonstration systématique des principaux théorèmes et résultats

auxquels elles donnèrent lieu. En réunissant ainsi dans un même volume tous les acquis formels qui contribuèrent à la constitution du paradigme catégoriel issu du calcul syntaxique de Lambek (1958), Richard Moot et Christian Retoré nous livrent un véritable ouvrage de référence qui, bien que n'étant pas de lecture aisée, s'avère être un outil d'autant plus précieux qu'il contient des bibliographies très exhaustives de chaque thème étudié ; même s'il faut toutefois déplorer que les pages d'exercices, systématiquement proposées à la fin de chaque chapitre, ne soient pas accompagnées de leurs corrigés.

Ainsi les deux premiers chapitres consacrés au modèle catégoriel initial argumentent que l'intérêt et l'originalité de l'approche catégorielle résident tout d'abord dans sa capacité à pouvoir exprimer la grammaire en ne recourant qu'à un formalisme purement logico-mathématique qui offrira, comme le met en évidence le chapitre 3, l'avantage de pouvoir calculer la signification des phrases en connectant la syntaxe avec la grammaire de Montague ou la sémantique dynamique de la théorie des représentations du discours (DRT). Mais la seconde propriété remarquable des grammaires catégorielles est qu'elles mettent également en jeu un formalisme lexicalisé qui induit la possibilité de définir un ensemble très restreint de règles universelles en conférant au lexique, par les biais des catégories (dénommées également « types »), la tâche de contenir toutes les informations spécifiques à chaque langue naturelle. Il en résulte – et ceci est l'objet de ce livre – qu'il devient possible de prendre pour objet d'étude en lui-même les propriétés formelles de la grammaire universelle.

Dans cette perspective, un des principaux thèmes récurrents dont traite cet ouvrage porte sur les diverses solutions qui furent apportées pour remédier à la limitation des capacités expressives des grammaires catégorielles, que C. Gaifman et N. Chomsky avaient déjà dénoncée, mais qui ne sera véritablement démontrée que dans les années 90 grâce au théorème de M. Pentus établissant que les grammaires de Lambek ne génèrent que des langages hors contexte, qui par conséquent ne contiennent pas le même degré de complexité structurelle que celui mis en jeu par les langues naturelles. Or, malgré l'intérêt, dans le cadre de cette problématique, des divers résultats formels apportés, le lecteur se questionne cependant sur l'absence de commentaires portant sur les difficultés théoriques que Lambek dénonça lorsqu'il élabora son dernier modèle catégoriel en termes de prégroupe. En effet, en étant obligé de stipuler, d'une part, de nouvelles règles de réduction *ad hoc* entre types et de spécifier, d'autre part, des métarègles autorisant des changements de type en fonction du contexte (pour, par exemple, rendre compte respectivement des propositions relatives et de l'ordre des mots dans les subordonnées allemandes), Lambek précisa que l'introduction de telles règles mettait en péril le modèle algébrique lui-même, puisqu'elles ont pour conséquence que : « les types ne forment plus une *hiérarchie* définie de manière inductive et [que] nous ne pouvons plus soutenir avoir ramené l'entièreté de la grammaire dans le dictionnaire, car les règles de réduction ont le statut de règles grammaticales additionnelles ». Or, cette même question du statut des « règles » (et des « métarègles ») se retrouve dans la notion de

« postulat » que M. Moortgat définit au sein de sa logique multimodale des types pour augmenter l'expressivité de son modèle et, bien que le chapitre 5 de Moot et Retoré nous propose une analyse très détaillée et fort précieuse de cette théorie complexe, des réflexions de type épistémologique auraient cependant été les bienvenues pour permettre au lecteur de mieux cerner les enjeux théoriques nécessairement induits par le choix d'un nouveau cadre formel.

Cependant cette critique ne doit pas minorer qu'une des forces de ce livre est qu'il établit sans cesse des passerelles entre les divers formalismes mis en œuvre par les grammaires catégorielles afin soit d'élucider, dans une optique pédagogique, la relation de transposition possible qu'ils entretiennent entre eux (cas du calcul des séquents et de la déduction naturelle, ou de la correspondance pouvant respectivement être établie entre, d'une part, le calcul de Lambek et certaines variantes de la logique linéaire multiplicative, et d'autre part, les réseaux de preuves et le calcul des séquents), soit pour argumenter de la supériorité de l'un de ces formalismes pour mettre à plat l'organisation syntaxique des phrases. Ainsi le chapitre 6 défend l'idée que les réseaux de preuves constituent d'excellents outils permettant d'analyser de façon déductive les structures syntaxiques des phrases en résolvant les cas de « fausses ambiguïtés ». Le chapitre 7, qui clôt le livre, offre un prolongement de cette thèse en appliquant les réseaux de preuves au calcul non associatif de Lambek et à ses extensions multimodales. L'intérêt d'user d'une telle démarche est alors mis en évidence grâce à la présentation du parseur interactif Grail et de l'illustration, à partir de quelques exemples en français issus du corpus arboré de l'université Paris 7, de la méthode utilisée pour déterminer les structures sémantiques de textes ne relevant pas de cette base de données. Ce faisant, cet ouvrage nous livre – après avoir retracé dans ses précédents chapitres les avancées formelles ayant enfin permis l'analyse des faits empiriques jugés problématiques pour l'approche catégorielle – un aperçu non négligeable du champ d'application actuel des formalismes catégoriels à la linguistique computationnelle.

Manfred STEDE. Discourse Processing. *Morgan & Claypool publishers*. 2011. 155 pages. ISBN 9781608457342.

Lu par **Richard MOOT** et **Christian RETORÉ**

CNRS, LaBRI / Université de Bordeaux, LaBRI & IRIT

Le livre de Manfred Stede s'adresse aux étudiants et aux chercheurs en traitement automatique des langues ainsi qu'à ceux des domaines voisins : linguistes, chercheurs en recherche d'information (!). Cet ouvrage synthétique propose au lecteur un panorama clair et concis des techniques totalement automatisées d'analyse du discours — un titre plus explicite serait Automated discourse analysis.

Selon l'auteur, ladite analyse du discours comporte trois questions essentielles qu'il développe dans les trois chapitres principaux, et qui sont toutes traitées par des techniques statistiques de *machine learning* :

- segmentation en parties homogènes d'un texte ;
- calcul des chaînes anaphoriques (qui aident à la segmentation) ;
- recherche des unités discursives élémentaires (EDU) et des relations de discours qui les lient.

Une rapide introduction de sept pages explique en quoi et pourquoi les trois questions choisies par l'auteur sont cruciales dans l'analyse automatique du discours. Le livre commence sur la juste observation que la différence majeure entre une suite de phrases et un discours réside dans la cohérence et l'organisation du discours. Chacune des trois questions développées dans ce livre contribue à dégager cette structure discursive, laquelle s'avère fort utile à la recherche d'information, au résumé automatique, aux études d'opinions ou à la réponse automatique à des questions (*question answering*). L'auteur précise aussi qu'il n'abordera que des textes écrits à une seule voix, comme les articles de presse, et qu'il ne traitera ni de données orales, ni de dialogues, ni d'interviews transcrites, etc.

Le chapitre 2 porte sur la segmentation du texte, c'est-à-dire sur sa découpe en unités cohérentes. À cette fin, on peut utiliser la structure du texte, surtout lorsqu'il s'agit d'un document structuré (notices, critiques de films, etc.). La découpe en thèmes (*topics*) qui sont généralement des unités assez grandes est une tâche de classification, où, pour être efficace, on suppose que chaque phrase est dans un unique segment. Cette restriction est un peu abusive, puisque certaines phrases peuvent participer à deux segments, ce qui occasionne des difficultés d'évaluation. En effet, comme l'échantillon de référence est aussi annoté avec un seul thème par phrase, l'évaluation de la classification peut être mauvaise alors que la classification de la phrase commune aux deux segments est juste. La distribution des mots comme marque d'unité de segments discursifs est ensuite discutée. Celle-ci peut être judicieusement complétée par l'étude des anaphores et des connecteurs discursifs, qui font respectivement l'objet des deux chapitres suivants.

Le chapitre suivant traite effectivement des coréférences et des chaînes anaphoriques qui sont liées à la notion de « thème » discuté ci-dessus, car la référence est souvent un élément saillant du thème. L'idée est de segmenter le texte à la frontière de deux phrases lorsque peu de chaînes anaphoriques franchissent cette frontière. Bien sûr, les pronoms sont un cas d'anaphores, mais majoritairement il s'agit d'anaphores nominales (appelées anaphores associatives en français : *le président de la République, le chef de l'État, le locataire de l'Élysée…).* Une partie importante de ce chapitre porte sur l'annotation qui convient à ce type de tâche, car c'est une question assez délicate. Par exemple, on doit décider si l'anaphore reprend seulement la tête d'un groupe nominal ou si elle reprend aussi celle des compléments de noms et des adjoints, ce qui posera un problème en présence de relatives selon qu'elles sont appositives ou restrictives. On notera que cet écueil

pourrait être évité grâce à l'utilisation de variables, comme celles que manipule la sémantique formelle standard.

Le chapitre 4 présente la théorie des structures rhétoriques (RST) qui décrit la structure discursive par des relations entre noyau(x) et satellite(s). Celle-ci permet de représenter la structure discursive d'un texte. C'est à notre avis un choix moins judicieux que la théorie des représentations discursives segmentées (S-DRT) d'Asher et Lascarides. En effet, la RST peine à définir nettement la structure discursive élémentaire (EDU) alors qu'un système logique permet de la définir comme une proposition au sens logique, ce qui est nettement plus clair. L'auteur discute des paramètres des structures rhétoriques. Faut-il autoriser des relations entre des noyaux multiples et des satellites multiples ? Des raisons d'efficacité informatique conduisent l'auteur à recommander des structures rhétoriques binaires, avec des relations entre un unique noyau et un unique satellite ce qui conduit à une structure d'arbre binaire. Cette limite pose problème. En cas de relation entre cause conséquence, qui est le noyau ? Et en cas d'élaboration narrative, comment réduire la suite d'événements à une seule EDU ? Un autre paramètre est discuté, il s'agit du degré d'imbrication des structures rhétoriques.. Une structure complexe pouvant elle-même être considérée comme une structure élémentaire, à quel degré d'imbrication récursive convient-il de se limiter ?

Là encore, l'auteur suggère de se limiter à des structures relativement plates, cette fois en raison de l'inintelligibilité des structures trop imbriquées.

Le livre discute ensuite des connecteurs discursifs qui permettent de trouver les relations discursives entre EDU. Comme le dit l'auteur, ils sont assez difficiles à déceler et à interpréter : il peut y avoir ambiguïté entre une préposition et un connecteur, un même connecteur peut exprimer des relations discursives différentes, et surtout lesdits connecteurs sont souvent absents (*Paul est tombé. Jacques l'a poussé.)* Les chercheurs francophones peuvent pousser un cocorico, car les travaux d'annotation discursive du projet ANNODIS sont bien cités.

Une très brève conclusion (2 pages) propose deux pistes pour améliorer l'efficacité des méthodes proposées : d'une part l'utilisation simultanée de différents indices de rupture, d'autre part, en spécialisant les méthodes au genre du texte et au type de discours. Le genre d'un texte est sa catégorie formelle (article de journal, notice, critique de film, etc.) et il y a une trentaine de catégories, tandis que le type de discours, qui peut varier au cours d'un même texte, est à choisir parmi une petite dizaine de possibilités (argumentation, narration, description, etc.). Bien évidemment, la spécialisation devrait grandement améliorer les techniques d'analyse automatique du discours proposées dans cet ouvrage.

Ce petit livre de Manfred Stede est très bien écrit. Il prend certains partis pris qui facilitent la présentation, quitte à être parfois un peu schématique et restrictif. Par exemple, l'auteur a choisi de ne parler que de trois questions et de n'utiliser que des méthodes d'analyse statistique et de *machine learning.* Ces techniques sont certes majoritaires, surtout en pratique, mais sont-elles les seules pour autant ? N'y a-t-il

pas d'autres questions centrales dans l'analyse automatique du discours ? Plus ponctuellement, certaines restrictions relevées dans notre résumé des chapitres sont discutables. Un autre bémol, sans doute dû à la taille du livre, est le manque de détails pour celui qui voudrait effectivement développer ces méthodes sur machine. On notera, cependant, que l'ouvrage fournit des références bibliographiques pour découvrir d'autres questions et d'autres approches, ainsi que pour approfondir et programmer les techniques qui y sont décrites.

En fait, ce livre répond parfaitement à la spécification de la collection « *Synthesis lectures on human languages technologies* » de Morgan & Claypool : c'est une présentation concise et originale d'un sujet important de recherche et développement, et, dans cette optique, ce livre est quasi parfait, nous ne pouvons qu'en recommander la lecture.