Notes de lecture

Rubrique préparée par Denis Maurel

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

Raoul BLIN, Introduction à la linguistique formelle, *Hermès-Lavoisier*, 2009, 221 pages, ISBN 978-2-7462-2010-2.

Lu par **Guy PERRIER**

Loria, université Nancy 2

Le livre présente un ensemble d'outils mathématiques couramment utilisés pour formaliser la syntaxe et la sémantique des langues.

Cet ouvrage répond à un besoin pour tous ceux qui souhaitent s'initier à la linguistique formelle et qui n'ont aucun bagage mathématique particulier, si ce n'est celui qu'ils ont acquis en fréquentant le collège et éventuellement le lycée. Même si le livre se limite à la syntaxe et à la sémantique des langues, la formalisation de ces deux niveaux fait appel à des outils mathématiques qui ne sont pas toujours faciles à appréhender : les langages formels et les grammaires algébriques, la logique du premier ordre et le lambda-calcul, entre autres.

Le mérite de Raoul Blin est de les présenter de façon pédagogique en liant l'introduction des outils mathématiques à leur application à la linguistique. À ce propos, on ne peut que se réjouir du point de vue de l'auteur : la linguistique est une science expérimentale au même titre que la physique ou la biologie et on peut lui appliquer les méthodes de la science expérimentale, ce que ne manque pas de faire Raoul Blin. Comme fil conducteur des exemples qui jalonnent le livre, l'auteur a choisi la formalisation de la syntaxe et de la sémantique de la phrase copulative *groupe nominal – copule – groupe nominal*. L'application choisie semble pertinente, n'étant ni trop pauvre, ni trop complexe. Il est appréciable que les notions introduites soient accompagnées d'exemples et d'exercices, et que ces exercices soient corrigés. On peut seulement regretter que les exercices soient parfois trop élémentaires.

Du point de vue des notions couvertes, on y trouve les notions de base en théorie des ensembles et des langages formels (chapitre 2), les grammaires algébriques (chapitre 3), la logique propositionnelle et la logique du premier ordre (chapitre 4) et le lambda-calcul non typé (chapitre 5). On peut regretter que l'auteur s'étende un peu trop longuement sur des notions de base relativement simples en théorie des ensembles, de même que sur la logique des propositions, alors qu'il fait totalement l'impasse sur des concepts pourtant très utilisés en linguistique formelle : les

graphes, en particulier les *graphes acycliques orientés* (utilisés notamment pour définir formellement les structures de traits et l'unification) et les *arbres*. Enfin, présenter le lambda-calcul sans présenter le lambda-calcul typé quand on cherche à formaliser la sémantique, c'est s'arrêter au milieu du gué : les types sont essentiels dans la sémantique formelle vue à travers le lambda-calcul.

Le livre se veut aussi le plus neutre possible par rapport aux théories linguistiques, mais toujours est-il qu'il se place quand même dans la vision syntagmatique de la syntaxe et dans l'approche à la Montague de la sémantique, ignorant totalement les grammaires de dépendances qui occupent pourtant un terrain de plus en plus grand en linguistique formelle.

Les deux derniers chapitres sont plus prospectifs. Le chapitre 6 montre toute la difficulté de passer de la sémantique de la phrase à la pragmatique où la représentation du contexte se heurte à plusieurs difficultés (masse des connaissances à représenter, relativité des connaissances par rapport aux univers mentaux des différents locuteurs…).

Le chapitre 7 propose un formalisme syntaxique fondé sur la logique du premier ordre, qui se veut un dépassement des limites rencontrées par les grammaires algébriques. Ce chapitre n'est pas du tout convaincant et il ignore complètement l'état de l'art. Le formalisme proposé ressemble aux DCG (*Definite Clause Grammars*) proposées en 1980 par Pereira et Warren mais depuis, les DCG ont été supplantées par de nouveaux formalismes tels TAG, LFG ou HPSG. D'une façon générale, il est regrettable que la bibliographie ne fasse référence qu'à des travaux déjà anciens et ignore tous les développements plus récents en syntaxe et en sémantique formelle.

Examinons maintenant plus en détail les chapitres principaux (chapitres 2 à 4).

Chapitre 2. Notions de base

Ce chapitre vise à introduire des notions de base en théorie des ensembles : les opérations d'union et d'intersection, la relation d'inclusion, le produit cartésien, les relations et fonctions.

L'introduction d'un pseudo lambda-calcul en plein milieu du chapitre pour représenter l'aspect compositionnel de la sémantique de la phrase est plutôt mal venue. Premièrement, la présentation rate l'essentiel qui est de montrer qu'il s'agit de représenter des fonctions. Par exemple, l'auteur nous donne le lambda terme représentant la sémantique du déterminant, sans jamais nous dire qu'il s'agit d'une fonction qui prend en argument une propriété (ou un ensemble d'entités) pour retourner un ensemble d'entités. La cohésion du chapitre en aurait été renforcée si l'auteur avait utilisé le concept de fonction, tel qu'il est défini dans la théorie des ensembles à la fin du chapitre, pour modéliser la composition sémantique. En plus, cela aurait permis de mieux motiver l'introduction du lambda-calcul au chapitre 5, celui-ci apparaissant alors comme un moyen élégant de calculer avec des fonctions.

De la même façon, pour que le chapitre reste focalisé sur l'introduction de notions de base de théorie des ensembles, il aurait mieux valu renvoyer la présentation des langages formels au chapitre suivant en préliminaire aux grammaires algébriques. Au lieu de cela, celle-ci s'enchevêtre avec les notions de théorie des ensembles, ce qui donne au chapitre une apparence de fourre-tout.

Par ailleurs, l'auteur passe trop rapidement sur la notion de *définition récursive* qui peut apparaître au lecteur non averti comme une arnaque dans la mesure où la définition récursive d'une notion utilise cette même notion. Bien entendu, en dire un peu plus nous amène à parler de la notion de point fixe, mais ne faut-il pas en passer par là pour convaincre le lecteur intrigué que l'on n'est pas en train de faire du bricolage ?

Chapitre 3. Syntaxe et grammaires formelles

Ce chapitre est particulièrement réussi, car il montre bien comment dans une démarche scientifique expérimentale, on en arrive à partir de corpus d'observation à bâtir tout d'abord la notion de syntagme et de catégorie syntagmatique en utilisant le test de commutativité. Ensuite, à partir de la décomposition immédiate d'un syntagme en constituants immédiats, on en vient à la notion de règle d'une grammaire syntagmatique. À la fin du chapitre, est même montré le lien entre composition syntaxique et composition sémantique, à travers l'exemple récurrent de la phrase copulative.

La plupart des notions relatives aux grammaires algébriques sont abordées : dérivation, arbre de dérivation, pouvoir génératif, équivalence faible des grammaires… Il est seulement dommage que la définition de *grammaire récursive* soit incorrecte : une grammaire récursive ne contient pas nécessairement de règle récursive car la récursivité se définit au niveau des dérivations et non des règles.

Chapitre 4. Sémantique et logique formelles

Ce chapitre est essentiellement consacré à la présentation de la logique des propositions et de la logique des prédicats à la fois sous l'angle de la théorie de la démonstration et sous l'angle de la théorie des modèles.

Il est dommage que l'auteur ait choisi une présentation de la logique à la Hilbert, pour un livre qui s'adresse à des non-spécialistes. Les axiomes ne sont pas du tout intuitifs, pas plus que les démonstrations. Au contraire, une présentation utilisant le cadre de la *déduction naturelle*, avec une règle d'élimination et une règle d'introduction pour chaque connecteur logique, aurait été beaucoup plus abordable. La déduction naturelle rend beaucoup plus compréhensible le raisonnement par contradiction, le raisonnement hypothétique ou le raisonnement par cas.

La présentation de la méthode de résolution, qui est essentiellement motivée par la démonstration automatique, n'a pas sa place, à mon avis, dans un livre qui s'intéresse à la formalisation linguistique. Elle relève plutôt du traitement automatique des langues.

Conclusion

Même si les notions abordées ne sont pas toutes présentées de la façon que l'on aurait souhaité, ce livre répond à un besoin important chez tous ceux qui veulent s'initier à la linguistique formelle et espérons qu'il les aidera à s'approprier les outils mathématiques nécessaires pour cela.

Iva NOVAKOVA, Agnès TUTIN, Le Lexique des émotions, *Éditions littéraires et linguistiques de l'université de Grenoble*, 2009, 349 pages, ISBN 978-2-84310-149-6.

Lu par **Maryvonne HOLZEM**

Laboratoires LiDiFra EA 4305 & LITIS EA 4108, Université de Rouen

---

Cet ouvrage, dont une partie des seize contributions provient d'un colloque tenu en 2007 à Grenoble, jette un éclairage novateur et varié sur un lexique souvent difficile d'accès, mais ô combien important en neurosciences, psychologie cognitive ou philosophie. En introduisant l'ouvrage, Iva Novakova et Agnès Tutin rappellent l'indissociabilité entre rationalité et affect. Elles s'inscrivent dans le prolongement de la grammaire des sentiments d'Antoinette Balibar-Mrabti (1995) et portent un intérêt tout particulier à la façon dont la combinatoire syntaxique et lexicale des noms d'émotions peut aider à comprendre la structuration de ce champ sémantique. Le lexique des émotions (les entités uniquement nommées ici, comme : joie, amour, bonheur, colère, peur, angoisse,…) est étudié dans des langues variées (français, espagnol, russe, polonais, grec) ce qui en souligne l'intérêt. Plusieurs approches linguistiques, d'inspiration structurale ou cognitive sont évoquées, toutes accordent une place essentielle à la combinatoire linguistique. Si les études proposées dans cet ouvrage abordent des enjeux théoriques essentiels comme la construction du sens à travers les associations des lexies exprimant des affects ou la structuration du champ sémantique des émotions, elles débouchent aussi sur des applications utiles pour la linguistique, comme l'enseignement structuré de la phraséologie en français langue étrangère (FLE), la classification automatique des unités lexicales en traitement automatique du langage (TAL), ou un traitement plus systématique en lexicographie.

Dans la première contribution, Gerda Haßler cherche à déterminer l'impact des définitions métalinguistiques d'Étienne Bonnot de Condillac (extraites du *Traité des sensations* de 1788) sur le sens actuel de ces mots. L'approche historique des occurrences de quelques noms met en lumière les relations de polarité, de variantes et d'hyperonymie dans lesquelles entrent ces noms. Par une étude comparée avec d'autres textes, cet article montre à quel point le sens d'un mot est fonction de l'histoire de ses interprétations au sein des textes et des contextes.

Peter Blumenthal jette les bases d'une cartographie du lexique des émotions facilitant la comparaison des systèmes de classification et de schématisation. L'auteur met en évidence un ensemble de propriétés sémantiques du champ lexical,

mais se garde bien de dresser des tendances collectives et prédictibles. Il émet notamment des doutes sur le fait que les hyponymes d'un même hyperonyme puissent avoir le même comportement combinatoire par héritage. Remarque qui nous semble très pertinente à l'heure des fusions multiterminologiques.

Agnès Tutin et Iva Novakova présentent une étude centrée sur les déterminants à partir d'un large corpus issu de la base de données Frantext. Certaines corrélations, jusqu'ici méconnues, entre le nom d'affect et sa distribution sont soulignées. Plusieurs distributions comptables et massives sont présentées, tout comme sont prises en compte la dimension aspectuelle et l'observation des cooccurrences lexicales. Cette étude bien documentée confirme la variabilité de la détermination des noms d'affect qui ne constituent pas une classe homogène. Cette question appelle à la prise en compte d'autres dimensions pour comprendre le fonctionnement de la détermination.

La contribution suivante se donne pour but l'élaboration d'une grammaire des sentiments à partir d'une description de 85 noms de sentiments en grec. Partant des propriétés actancielles, elle dégage les grands traits d'une combinatoire lexicale ouvrant sur une classification syntaxique fine. Il est seulement à regretter que les auteurs n'aient pas mieux circonscrits et datés leur corpus d'analyse.

Le chapitre d'Angels Catena et Effi Lampou est orienté sur la traduction. Il offre une typologie des prédicats d'affect en espagnol et grec pour mieux cerner les problèmes soulevés par l'attribution d'équivalents entre ces deux langues.

C'est à partir de la distribution des adverbes *parfaitement* et c*omplètement* que Danielle Leeman émettra des hypothèses quant aux adverbes, dits de complétude, exprimant un état ponctuel lorsqu'ils cooccurrent avec les acceptions de *malade* dans ses emplois de perturbation mentale ou affective.

Le troisième volet de cet ouvrage est consacré à des études distinctives. Tout d'abord d'un point de vue aspectuel avec Houda Ounis autour des notions de *coup de foudre* et *amour*. Au terme d'une analyse compositionnelle, elle démontre que la distinction entre ponctuel/duratif est déterminante, tout comme celle entre caractère exogène (*coup de* comme venant de l'extérieur)/endogène (*amour* comme émanant d'un « intérieur »).

C'est aux N_sent (noms de sentiments dans la typologie des noms d'affect) incarnés par des prototypes nominaux, qui en sont les « meilleurs exemplaires » (*cf*. Tutin) que s'intéressera Elena Melnikova. Elle présentera en détail son corpus d'étude (corpus parallèle de traduction russe/français de textes des XIX^{e} et XX^{e} siècles) à partir duquel elle a extrait un patron syntaxique (Vsup (verbe support) + N_sent.). L'auteur conclut alors à une distribution des N + sent. très diversifiée entre le français et le russe.

Deux contributions sont consacrées à la structuration du champ sémantique d'un point de vue métaphorique. Analyse de la *tristesse* avec Anna Krzyzanowska et de la *colère* avec Magdalena Augustyn et Ekaterina Bouchoueva. Dans les deux cas, c'est

l'étude de la structure actancielle entrant dans la combinatoire syntaxique et sémantique qui sera mise en évidence. Si les auteurs s'accordent sur les régularités au niveau des collocations, des divergences, entre le français d'un côté, et le russe et le polonais, de l'autre, sont mises en évidence au niveau des structures actancielles issues de l'approche contrastive entre ces trois langues dans la seconde étude.

L'ouvrage s'achève sur la phraséologie. Céline Vaguer propose une caractérisation des locutions prépositives incluant un nom d'émotion en français fondée sur une ressource lexicale jusqu'alors peu étudiée. Par l'élaboration de ressources multilingues l'auteur tente de façon assez convaincante de pallier les carences des approches informatiques.

Freiderikos Valetopoulos travaille sur l'interface syntaxe et lexique à travers les constructions verbales *avoir* et *être* suivies d'un nom de sentiment en grec moderne. Il met en évidence un continuum entre expressions libres et figées permettant de mieux comprendre la combinatoire lexicale et syntaxique des noms dits psychologiques.

Le dernier volet sera consacré aux applications linguistiques de la combinatoire des émotions. Le travail de Margarita Alonso Ramos porte sur la polysémie régulière des noms de sentiments. Il s'inscrit dans cadre dictionnairique des collocations de l'espagnol (DiCE). L'auteur propose d'étendre le codage de la lexicologie explicative et combinatoire de Mel'cuk en indiquant clairement la nature du second actant sémantique (cause ou objet : $N_{obj/caus}$) de façon a pouvoir dériver plus clairement le nom associé.

Sigrid Maurel, Paolo Curtoni et Luca Dini proposent une nouvelle méthode Sybille (approche hybride à la fois symbolique et statistique) de classification automatique capable d'améliorer la qualité d'un système d'extraction de sentiments à partir d'une démarche centrée sur les utilisateurs. Le corpus de cet article traite de textes d'opinion.

L'ouvrage s'achève par l'enseignement de la phraséologie en FLE, plaidant pour la construction d'un enseignement onomasiologique pour les expressions associées aux affects permettant aux apprenants (présentation ici d'un projet mené en milieu scolaire) de mieux comprendre les structures lexicales de la langue en apprenant à manipuler la variation.

En conclusion, cet ouvrage bien que ne présentant pas (sauf pour une contribution) de nouvelle méthode de TAL, démontre à quel point les apports logiciels (notamment dans le traitement des concordances) ont enrichi les études lexicologiques permettant une description fine et contrastive des langues. Il témoigne également de la vitalité des recherches menées dans la lignée de la lexicologie explicative et combinatoire d'Igor Mel'cuk. Nous regrettons cependant, que les entités nommées, évoquées ici, ne soient pas mieux attestées. Elles gagneraient, selon nous, à être considérées en fonction des genres textuels et des corpus d'où elles ont été extraites.

Jean-Pierre DESCLÉS, Florence LE PRIOL, Annotations automatiques et recherche d'information, *Hermès Lavoisier*, 2009, 330 pages, ISBN 978-2-7462-2226-7.

Lu par **Annie Tartier**

Retraitée, anciennement membre de l'équipe TALN du LINA (UMR 6241)

Ce livre est un ouvrage collectif, composé de douze chapitres, et structuré en deux parties. Le thème fédérateur est celui de l'annotation automatique et de ses usages en recherche d'information. La première partie présente la méthode d'exploration contextuelle implémentée dans la plate-forme EXCOM. La deuxième partie décrit d'autres méthodes d'accès aux informations.

Première partie – Méthodes d'accès aux informations par exploration contextuelle : EXCOM

Pour donner un bref panorama de cette première partie on peut retenir que le premier chapitre est un article de fond qui doit nécessairement être lu en premier lieu car il explique la méthode d'exploration contextuelle et en présente une implémentation. Les trois chapitres suivants relatent des applications, très différentes les unes des autres, et peuvent intéresser des chercheurs en TAL qui travaillent sur les sujets concernés. Le dernier chapitre, moins applicatif, apporte au lecteur des compétences linguistiques et cognitives utiles pour mettre en œuvre la méthode d'exploration contextuelle dans la résolution de la polysémie.

Après avoir mentionné l'importance de trouver l'information pertinente au moment opportun, le chapitre 1 rappelle le principe des moteurs de recherche actuels. Pour mieux cibler certaines informations, la *méthode d'exploration contextuelle*, fondée sur la notion de *point de vue de fouille* est mise en œuvre dans deux plates-formes EXCOM et MOCXE. EXCOM annote automatiquement des textes bruts à partir de marqueurs linguistiques exprimant des relations sémantiques liées au point de vue de fouille considéré (définition, causalité, etc.). Ensuite, MOCXE indexe les annotations et les utilise pour les recherches. La dernière partie de l'article compare cette stratégie avec d'autres qui utilisent des ressources linguistiques beaucoup plus lourdes.

Le chapitre 2 décrit trois applications d'annotation sémantique, utilisant EXCOM.
1- Contrairement aux applications courantes qui ignorent les références entre différents types de médias, la première application annote automatiquement les fragments textuels référençant les images, quelles que soient leurs positions relatives dans le document, à l'aide de marqueurs linguistiques dédiés à ce type de fouille.
2- Après un long passage sur les avantages et inconvénients de la bibliométrie, la deuxième application utilise la plate-forme EXCOM pour annoter et catégoriser

automatiquement les fragments annonçant les références, et présente une distribution des citations plus qualitative qu'une valeur numérique obtenue par simple comptage. 3- La troisième application propose une alternative au résumé automatique sous la forme d'une *fiche de synthèse* qui contient les informations liées aux centres d'intérêt de l'utilisateur, informations extraites à partir du point de vue de fouille de celui-ci.

L'écriture actuelle de la langue arabe omet de nombreux signes de vocalisation, générant ainsi beaucoup d'ambiguïtés. Le chapitre 3 présente quatre particules dont le rôle dans la phrase dépend des signes de vocalisation du groupe de lettres *alef-noun*. Après avoir décrit deux méthodes de désambiguïsation utilisées en TAL, l'article explique sa méthode d'exploration contextuelle s'appuyant sur des marqueurs, repérés par des linguistes, qui permettent de caractériser le rôle des particules.

Ce travail vise à enrichir la plate-forme EXCOM avec des ressources pour la langue coréenne et un point de vue de fouille sur le repérage des citations directes et indirectes dans des textes journalistiques. Le chapitre 4 recense et classe des indicateurs, typographiques et linguistiques, nécessaires pour repérer les citations directes et indirectes. Il en explique la difficulté par une assez grande différence, en coréen, entre la pratique et la norme, mais aussi parce que les marqueurs recensés peuvent être utilisés dans d'autres contextes que les citations. Des exemples sont proposés.

Le chapitre 5 s'appuie sur la similitude entre le processus de lecture humaine pour accéder à la signification d'un texte, et la méthode d'exploration contextuelle appliquée à la recherche de la signification pertinente de verbes polysémiques. La polysémie des verbes est expliquée par des combinaisons de primitives sémantico-cognitives qui sont déductibles de la structure du contexte, s'il est bien précisé. L'article explique comment trouver les indices pertinents pour l'exploration contextuelle et décrit cinq règles d'héritage qui s'appliquent dans des cas non standard. Un exemple détaillé sur les significations du verbe *avancer* est proposé.

Deuxième partie – Quelques autres méthodes d'accès aux informations : exemples et problèmes sémantiques

Contrairement à la première partie, celle-ci n'a pas de fil rouge. Les chapitres 6, 7 et 8 sont des descriptions de méthodes ou de systèmes d'annotation sémantique complètement ou partiellement automatique qui peuvent enrichir des travaux de domaines similaires. Le chapitre 9 s'éloigne de l'annotation automatique et cherche à mettre en place une grammaire universelle multilingue. Par opposition le chapitre 10 est très spécialisé sur le thème du mot en albanais. Les deux derniers chapitres, qui concernent la logique des questions et la théorie des actes de discours sont davantage centrés sur des notions linguistiques.

Après avoir présenté cinq systèmes d'annotation sémantique, Le chapitre 6 aborde le sien, dont l'objectif est d'enrichir un document avec une sorte de résumé

sur les événements relatés, pour offrir une perception rapide ne nécessitant pas la lecture intégrale du document. Le processus d'annotation automatique catégorise chaque phrase en *événementielle* ou non, en s'appuyant sur un ensemble d'apprentissage annoté par un expert. Les *phrases événementielles* sont regroupées grâce à une nouvelle mesure de similarité et servent à construire le résumé.

Le chapitre 7 décrit *Larsa*, un système de fouille de textes qui recherche des relations sémantiques correspondant aux questions les plus fréquentes. L'utilisateur choisit des questions en rapport avec des relations prédéfinies (causalité, but, etc.) et complète avec des mots-clés. Le système extrait des phrases exprimant ces relations grâce à des motifs qui correspondent à la relation sémantique visée. Une évaluation détaillée compare les résultats de *Larsa*, pour deux corpus, avec ceux d'une annotation manuelle.

Le chapitre 8 présente une méthode pour analyser et interpréter, à partir d'un texte, un concept représenté par une forme linguistique canonique. La démarche est assistée par des outils informatiques (concordancier, *clustering*), mais l'interprétation des résultats est laissée à l'expert. Une phase d'annotation manuelle marque de manière pertinente les segments rangés dans les clusters. Ces annotations sont utilisées pour produire l'analyse conceptuelle finale. Un exemple détaillé est donné sur un texte philosophique autour de la forme canonique *MIND*.

Pour se rapprocher d'applications indépendantes de la langue, le chapitre 9 veut montrer que les grammaires catégorielles sont bien adaptées à ces objectifs d'universalité, parce que leur mécanisme de base (opérateurs appliqués à des opérandes) est indépendant de la langue. Pour appuyer cette thèse une application progressive et détaillée des opérateurs et des règles de la *grammaire catégorielle combinatoire applicative* est présentée sur les traductions d'une phrase en trois langues (anglais, français, arabe) qui diffèrent par l'ordre des mots. Puis l'article aborde le problème complexe de la coordination en détaillant toutes les étapes d'analyse sur différents énoncés des trois langues considérées.

Le chapitre 10 constitue une étude linguistique, préliminaire à une implantation informatique, sur la reconnaissance automatique du mot dans la langue albanaise. Une première discussion porte sur l'impossibilité de définir le mot comme unité linguistique de base et sur l'opportunité d'introduire le morphème comme unité plus pertinente, mais plus difficile à identifier. La deuxième partie recense, dans un corpus de textes albanais, des formes typographiques semblables et analyse les cas qui conduisent à un ou à deux mots au sens linguistique du terme. Elle présente aussi des cas d'amalgame où un mot typographique comprend plusieurs unités de sens.

Le chapitre 11 cherche à présenter une approche générale pour modéliser des dialogues homme-machine et les implémenter dans les sites Web. Au premier niveau la logique des questions établit une typologie des énoncés interrogatifs. Comme certains problèmes sont insolubles sur des d'énoncés analysés isolément, l'analyse des questions est intégrée à la théorie, plus générale, des actes de discours. Le troisième niveau est celui de l'analyse des conversations, qui tient compte des

contextes dans lesquels se déroulent les questions et se fonde sur le fait que la plupart des conversations sont intelligentes et rationnelles.

Le chapitre 12 se place dans le cadre de la théorie des actes de discours. En pensant et en parlant l'homme accomplit des actes composés de forces illocutoires et de contenus propositionnels. Un système formel décrit les forces illocutoires de base, et détaille les opérations logiques s'appliquant à leurs composantes ainsi qu'aux contenus propositionnels. Ces éléments logiques sont ensuite mis en relation avec des éléments linguistiques permettant aux locuteurs de s'exprimer. L'article conclut que la théorie des actes de discours enrichit la tradition classique de la grammaire universelle, en utilisant les ressources de la logique contemporaine.

Pour conclure, dans cet ouvrage très varié tant dans les thèmes abordés que dans les types d'analyse, les chapitres peuvent être lus de manière indépendante à deux exceptions près : le premier chapitre doit être lu avant chaque chapitre de la première partie, des lecteurs non linguistes auront peut-être intérêt à lire le chapitre 12 avant le chapitre 11.

Joseph MARIANI, *Spoken Language Processing*, *Wiley*, 2009, 487 pages, ISBN 978-1-84821-031-8.

Lu par **Laurence LONGO**

Laboratoire LiLPa (EA 1339), Université de Strasbourg

Cet ouvrage aborde à la fois des aspects fondamentaux et expérimentaux sur le traitement automatique du langage parlé. Les différentes facettes traitées permettent d'appréhender les mécanismes mis en jeu pour automatiser la production et la perception de la parole, mais aussi pour synthétiser et comprendre le langage parlé, et ce, à partir des connaissances acquises sur la parole naturelle et celles issues d'autres domaines tels le traitement du signal, la linguistique computationnelle, la reconnaissance des formes, l'ergonomie et les modélisations stochastiques.

Le traitement automatique du langage parlé comprend l'analyse de la parole, son codage, sa synthèse (à partir du texte), sa reconnaissance (appliquée à un locuteur donné ou à une langue spécifique), sa compréhension (pour la transcrire, l'indexer ou bien l'utiliser en dialogue homme-machine), que ces activités se déroulent en milieu bruité ou non.

Le codage de la parole s'est développé avec l'évolution des systèmes modernes de télécommunication. À l'heure actuelle, le codage demeure toujours nécessaire en bande large, pour pallier les problèmes de couverture et de qualité de la voix rencontrés avec l'utilisation de la fibre optique. Pour un codeur de parole, trois aspects sont à prendre en compte : le débit, la qualité et la complexité. Il est difficile de mesurer de manière objective les dégradations engendrées par le codage. Le

codage a pour principe de ne pas transmettre ce qui est redondant ni ce que l'oreille ne perçoit pas. Le codage audiovisuel de la parole permet d'établir une complémentarité entre les informations auditives et les informations visuelles de la parole, réduisant de ce fait les bruits et autres pertes dus à la compression.

En synthèse de la parole, différents objectifs sont à atteindre pour élaborer un système : créer la matière acoustique (ou la couleur des sons), déterminer le contenu prosodique du message à prononcer (identifier la différence entre un fondamental montant et un fondamental descendant, par exemple sur la dernière syllabe dans « Il pleut. » ou « Il pleut ? »), éliminer l'ambiguïté entre l'émetteur et le récepteur du message parlé (en faisant intervenir la syntaxe et la sémantique).

Les traitements linguistiques, tels que le prétraitement du texte qui consiste à désambigüiser les acronymes, les entités nommées (noms propres de personnes, lieux et organisations, unités monétaires…), les numéros de téléphone ou d'autres abréviations, permettent de fournir au système de synthèse des textes qu'il est en mesure de « comprendre » (il s'agit souvent de transcrire sous forme littérale ces formes ambiguës). Un second traitement linguistique, la phonétisation, permet de faire correspondre la séquence de phonèmes adéquats à une prononciation de texte donnée. Cette opération doit pallier les phénomènes de coarticulation, de liaison et d'homographie propres à chaque langue. Diverses approches par règles ou lexiques permettent de résoudre ces ambigüités, mais il demeure encore le problème du traitement des noms propres. De son côté, l'analyse syntaxico-prosodique vise à définir la structure prosodique la plus pertinente et « naturelle » dans un message à lire. Les systèmes à base de règles effectuent une analyse syntaxique, de surface ou complète. Cependant, on constate que ce sont les modélisations de la structure de la prosodie par arbres de décision qui, en déterminant automatiquement la structure à partir d'exemples d'apprentissage, sont les plus pertinentes.

L'évaluation de la synthèse de la parole permet d'analyser la qualité de la voix de synthèse. Il s'agit d'évaluer les progrès mais aussi de pointer les défauts du synthétiseur. La plupart des tests reposent sur un jugement humain appréciant la qualité de la synthèse : intelligibilité, compréhension, débit, netteté, etc. Sont aussi évalués de manière indépendante les divers modules de synthèse, soit par des humains, soit, de manière semi-automatique, par les logiciels de tests développés au cours des diverses campagnes d'évaluation en télécommunication. La synthèse de la parole à partir de textes est opérationnelle dans les services de télécommunication, mais, à la différence de la commande vocale, elle est réservée aux services où elle se révèle indispensable pour passer de l'écrit à l'oral.

Pour compenser en partie les dégradations acoustiques d'un message parlé en situation bruitée, nous avons tendance à regarder le visage de notre interlocuteur. C'est cet aspect bimodal intrinsèque de la parole que les applications ayant pour objet des visages de synthèse, ou avatars, tendent à mettre au point pour améliorer l'intelligibilité des synthétiseurs de parole. Pour ce faire, les mouvements faciaux imposés au visage synthétique doivent être cohérents avec l'onde acoustique qu'ils

doivent produire (sinon l'auditeur serait trompé). L'animation des visages reste encore à l'heure actuelle une tâche complexe, car la synchronisation labiale avec le signal de parole nécessite une parfaite connaissance des processus de production et de perception de la parole.

En reconnaissance de la parole, la principale difficulté de l'analyse du langage naturel repose sur l'impossibilité, pour un système, de détenir une grammaire capable de générer de manière exhaustive les phrases du langage naturel et uniquement celles-ci. Les principaux composants à la base d'un système de reconnaissance de parole sont les corpus de paroles et de textes (le modèle de langage), le dictionnaire des prononciations acoustiques et le modèle acoustique. Ce faisant, en reconnaissance automatique de la parole, ce sont des grammaires surproductrices qui sont utilisées car elles offrent la possibilité d'être représentées par des automates stochastiques à états finis. Dans la plupart des systèmes de reconnaissance, cette dernière est réalisée par un décodage probabiliste qui va choisir l'événement linguistique de plus haute probabilité correspondant aux données observées. Un modèle de langage robuste permettrait d'augmenter les taux de reconnaissance en milieu bruité (qui est de l'ordre de 3 % à l'heure actuelle) ou en parole conversationnelle. Bien que la reconnaissance de la parole ne demeure pas encore un problème résolu, la forte quantité de données disponibles rend déjà possible la portabilité d'un système vers une nouvelle application (communication homme-machine) et même vers une nouvelle langue. Les systèmes utilisant les technologies vocales actuellement commercialisées vont de la dictée vocale à des serveurs vocaux, en passant par des systèmes de commandes ou de dialogues restreints. Cependant, l'utilisation courante des technologies vocales ne sera possible qu'à partir du moment où les performances des systèmes seront élevées, leur utilisation aisée et qu'ils répondront aux besoins réels des utilisateurs.

À l'inverse de la reconnaissance de la parole, pour la reconnaissance du locuteur l'attention se focalise sur les spécificités du locuteur pour caractériser sa voix sur un énoncé donné. La signature vocale regroupe des différences morphologiques, physiologiques et socioculturelles auxquelles se rajoutent d'autres variabilités interlocuteurs (état de santé, état émotionnel, intentions…). Parmi les diverses tâches à accomplir par les systèmes de reconnaissance du locuteur, les tâches centrales dans ce domaine sont l'identification du locuteur, dans un ensemble fermé de locuteurs référencés, et la vérification du locuteur. Pour ces deux tâches, d'excellentes performances sont obtenues dans des conditions idéales artificielles, mais elles se dégradent en situation réelle, ce qui nécessite la mise en œuvre d'approches spécifiques pour améliorer la robustesse des systèmes. En phase d'apprentissage les caractéristiques d'un locuteur et la modélisation probabiliste permettent d'établir un modèle du locuteur (un modèle de la distribution des énoncés qu'il produit). Pour la prise de décision dans la reconnaissance du locuteur, les modules de décision s'appuient sur les plus proches voisins, pour l'identification, et sur la comparaison à un seuil pour la vérification, après une phase d'ajustement du score. De par une bonne acceptabilité de la part de l'utilisateur, les applications de la

reconnaissance du locuteur sont nombreuses : transactions bancaires, télécommunications, jeux… Elles sont souvent couplées à d'autres technologies vocales pour réduire la fraude, par exemple.

Spoken Language Processing est un ouvrage qui mêle avec aisance les aspects théoriques et pratiques. À chaque fois, le lecteur est mis en garde à propos des adaptations et/ou des ajouts nécessaires au passage du modèle théorique à sa mise en application réelle. C'est un véritable guide du traitement automatique du langage parlé que l'on ne peut que vivement recommander.

ChengXiang ZHAI, *Statistical Language Models for Information Retrieval*, *Morgan & Claypool Publishers*, 2009, 125 pages, ISBN 9781598295900.

Lu par **Christian MAUCERI**

IBM-ILOG

Ce livre est une revue systématique des différents modèles statistiques de la langue utilisés en recherche documentaire (information retrieval*). Le problème essentiel de la recherche documentaire (RD) est de classer par ordre de pertinence les documents retrouvés dans un fonds documentaire par une requête. L'auteur résume un vaste corpus d'articles traitant ce problème d'un point de vue probabiliste qui rompt avec le traditionnel modèle vectoriel de Gerard Salton. Il met en avant la meilleure cohérence théorique de cette approche qui reste cependant aussi efficace que le modèle traditionnel en termes de performances.*

Le modèle vectoriel de Gerard Salton a longtemps dominé la littérature scientifique consacrée à la recherche documentaire. Depuis quelques années, une approche probabiliste issue des modèles statistiques de la langue prend une ampleur grandissante. Si elle reprend l'extrême simplification du document « sac de mots » où tout ordre linéaire est perdu, elle tente de mieux formaliser le problème : c'est cet effort que M. Zhai s'attache à synthétiser.

Il commence par rappeler, en introduction, les grandes lignes du modèle traditionnel et des modèles statistiques de la langue. Ceci lui permet d'introduire au second chapitre une axiomatique de la RD dans un cadre probabiliste, il introduit, en particulier, les importantes questions de minimisation du risque et de pertinence en RD dans une optique bayésienne.

Le chapitre 3 introduit la notion de vraisemblance d'une requête étant donné un modèle de document comme mesure de l'importance d'un document pour une requête. C'est ici que le lien avec les modèles de statistiques de la langue est clairement mis en lumière et où est introduite la notion de pertinence : le modèle du document prenant, bien entendu, en compte sa pertinence pour la requête au travers d'exemples préalables ; par exemple, les sélections de documents par des utilisateurs lors de précédentes requêtes. Trois modèles statistiques sont présentés et comparés

dans ce cadre : modèle multinomial, modèle de Bernoulli et modèle de Poisson. Dans ce même chapitre, l'importante question du lissage est abordée permettant d'évaluer la probabilité d'un mot dans une requête quand ce mot n'est jamais apparu dans des requêtes précédentes. Plusieurs techniques traditionnelles sont présentées : en particulier, le lissage de Jelinek et Mercer fondé sur le maximum de vraisemblance et celui, dans un cadre bayésien cette fois, fondé sur la probabilité préalable de l'hypothèse par une distribution de Dirichlet. Finalement, il est montré que certaines relations peuvent être établies entre le lissage et la fréquence de document inverse, notion centrale dans le modèle vectoriel traditionnel. Le chapitre 4 est, en fait, une prolongation du chapitre 3 sur des questions assez théoriques.

Le chapitre 5, introduit tout d'abord la notion de divergence de Kullback-Leibler afin de prendre en compte la notion de modèle de requête au même titre qu'au chapitre 3 était introduite la notion de modèle de document. Ainsi, le score d'un document en fonction d'une requête est assimilé à la divergence de Kullback-Leibler entre le modèle du document et le modèle de la requête. En particulier, lorsque le modèle de la requête est assimilé à la distribution empirique de ses mots, le score est équivalent à celui donné par la méthode décrite au chapitre 3. Le reste du chapitre est naturellement consacré à l'estimation de modèles de requêtes. Une première famille d'estimateurs se fonde sur une modélisation *a priori* de documents de rétroaction positive (les documents sélectionnés à la suite d'une requête). Une seconde approche, fondée sur la cooccurrence de mots dans les documents, modélise un processus de Markov où l'on passe d'un mot à un document puis d'un document à un mot idéalisant le parcours d'un internaute rebondissant d'un document à un autre en fonctions des mots qu'il y rencontre. L'auteur clôt ce chapitre sur un modèle de pertinence essentiellement fondé sur le rapport de la probabilité de retrouver un document pertinent connaissant la requête et sur la probabilité de retrouver un document non pertinent connaissant la requête. Il est enfin fait mention de deux approches : l'une concernant les requêtes structurées et l'autre la rétroaction négative de pertinence.

Le chapitre 6, traite de différentes tâches, citons pour mémoire : la recherche multilingue, la recherche distribuée, la recherche d'experts, la recherche dépendant du contexte et la recherche de passages. Il est dommage que ces thèmes très importants et différents soient regroupés pêle-mêle dans un même chapitre.

Le septième et dernier chapitre s'attarde sur l'analyse de thèmes latents qui répond d'une certaine façon à la sémantique latente du modèle vectoriel. L'idée est de découvrir des thèmes latents modélisés comme des distributions de mots, représentant les documents. L'analyse sémantique latente probabiliste (PLSA pour *Probabilist Latent Semantic Analysis*) proposée par Hoffmann en 1999 est d'abord présentée, et ses limitations génératives mises en évidence. La répartition latente de Dirichlet (LDA pour *Latent Dirichlet Allocation*) pour générer un document de longueur donnée est décrite ainsi que ses relations avec la méthode précédente.

Différentes extensions de cette notion de thèmes latents sont ensuite brièvement exposées.

Classiquement l'auteur conclut par une revue des avantages des modèles probabilistes sur les modèles traditionnels et une série de futures voies de recherche.

On peut regretter que les sujets les plus intéressants soient un peu trop rapidement expédiés dans les deux derniers chapitres, mais l'auteur a visiblement été guidé par un souci didactique, le poussant à bien formuler le cadre probabiliste sous-jacent. Le chapitre 3 est de ce point de vue particulièrement réussi. Mais ce qui fait avant tout l'intérêt de l'ouvrage est la très riche bibliographie qui l'accompagne, l'essentiel de la thématique, parfois ardue étant clairement présentée en une centaine de pages. Un livre à garder à portée de main.

Béatrice LAMIROY (coordinatrice), Jean-René KLEIN, Jacques LABELLE, Christian LECLÈRE, Annie MEUNIER et Corinne ROSSARI. Les expressions verbales figées de la francophonie, Belgique, France, Québec et Suisse, *Éditions Ophrys*, 2010, 163 pages, ISBN 978-2-7080-1238-7.

Lu par **Laurence DANLOS**

Université Paris Diderot, ALPAGE (UMR INRIA-Rocquencourt/Paris Diderot)

Les auteurs ont repris le lexique des 40 000 expressions verbales figées du français de France, développé (mais non publié) par Maurice Gross, en l'étendant aux variantes belges, québécoises et suisses. Ainsi l'expression coûter les yeux de la tête *s'emploie dans les quatre variétés du français,* coûter un os *uniquement en Belgique,* coûter une beurrée *uniquement au Québec,* coûter le lard du chat *uniquement en Suisse,* coûter bonbon *en France et en Suisse. Les auteurs, appartenant aux quatre pays des variantes francophones concernées, décrivent les fondements linguistiques qui sous-tendent la constitution de ce lexique et les résultats qu'ils en ont tirés.*

Le premier chapitre décrit les critères classiques (opacité référentielle, ruptures paradigmatiques…) permettant de qualifier une expression figée et rappelle que la notion de figement n'est pas binaire, mais relève d'une question de degré (par exemple *prendre une veste* est plus figée que *prendre un verre*).

Le second chapitre est consacré aux variétés du français et, après une brève introduction sur les données géographiques et historiques des quatre variétés concernées, donne les principes de classification utilisés qui s'organisent autour de 15 classes : BFQS regroupe les expressions qui sont utilisées dans les quatre variétés, B les belgicismes, F les francismes, Q les québécismes, S les helvétismes, les 11 classes restantes constituant les intersections possibles à deux ou trois éléments (ainsi *coûter bonbon* est classée en FS).

Le troisième chapitre étudie la morphosyntaxe des expressions verbales figées. Comme l'avait souligné Maurice Gross, on retrouve des irrégularités à tous les niveaux. Ainsi l'expression *avoir son permis de conduire dans une pochette surprise* (BFS) demande que le verbe *avoir* soit conjugué au passé composé ; la négation est obligatoire dans *ne pas se peigner avec un clou* (B). Contrairement à Maurice Gross, les auteurs n'ont pas recours aux symboles N_i pour les arguments libres et C_i pour les constantes : ils soulignent les parties figées. Par exemple, la phrase *Paul apporte de l'eau au moulin de Luc* relève de la structure N V N Prép [N Prép N] et non de la structure N_0 V C_1 Prép [C_2 Prép N]. Il est dommage que ce changement de notation ne soit pas justifié.

Le travail présenté dans le chapitre IV est tout à fait original par rapport à celui de Maurice Gross dans la mesure où les auteurs se sont attelés au sens des expressions figées : ils associent à chaque forme figée une forme libre de même sens. Ainsi l'expression *avaler des couleuvres* (BFS) a son sens décrit par 'être obligé d'accepter sans protester quelque chose de désagréable'. Ce chapitre décrit les problèmes délicats rencontrés pour attribuer un sens à une expression figée, problèmes qui sont amplifiés par l'étude simultanée des quatre variétés du français. Il faut avoir recours à la notion de « géosynonyme » : *en avoir ras le bol* (BFS) est un géosynonyme de *avoir son load* (Q), les deux signifiant 'ne plus pouvoir supporter' dans des variétés différentes du français. Il faut aussi avoir recours à la notion de « faux-ami » : *avoir de l'allure* signifie 'avoir de la distinction' dans les quatre variétés du français mais signifie aussi 'être tout à fait raisonnable' en québécois.

La dernière section du chapitre IV est consacrée aux expressions dont tous les éléments sont fixes et qui n'ont qu'un « sens pragmatique » dans la mesure où elles ne s'interprètent qu'en fonction de la situation d'énonciation dans laquelle elles sont prononcées. Ainsi, le sens de l'expression *On n'arrête pas le progrès* (BFQS) est noté 'formule souvent ironique concernant une nouveauté', celui de *Il va pleuvoir* (BFS) est noté 'formule pour indiquer que quelqu'un chante faux'.[1]

Si nous résumons, ce livre présente d'une part un résumé de l'étude lexicale et morphosyntaxique des expressions verbales figées effectuée par Maurice Gross – pour les lecteurs qui ne connaissent pas ces travaux sur le figement, c'est un incontournable , d'autre part, le livre étend cette étude dans deux directions : la prise en compte de quatre variétés du français et le sens des expressions figées, travail original et bien mené.

Ce livre est complété par une annexe qui donne un extrait du dictionnaire comportant 148 expressions figées, leur appartenance à une ou plusieurs variétés du

1. Doit-on en déduire que personne ne chante faux au Québec ou qu'il ne pleut jamais ou que l'on a la délicatesse de ne pas faire remarquer à quelqu'un qu'il chante faux ? Dit-on *Il va neiger* en de pareilles circonstances ?

français, leur forme syntaxique identifiant les parties constantes, leur sens décrit par une expression libre, un exemple de corpus, et, si besoin est, les expressions synonymes ou géosynonymes, ainsi que les variantes syntaxiques (*i.e. attraper/recevoir un cigare de qqn* vs *passer un cigare à qqn)*. On ne peut qu'amèrement regretter que cette annexe ne comporte que si peu d'entrées alors que, connaissant les auteurs, je sais qu'après une dizaine d'années de travail, ils ont pratiquement fini le dictionnaire pour toutes les expressions dont le verbe commence par la lettre « a » : soit environ 1 500 entrées dont le verbe est *avoir* et plus de 600 entrées dont le verbe n'est pas *avoir*. Ces dernières se trouvent sur un site Web qui n'est pas référencé dans l'ouvrage. On peut vraiment se demander pourquoi les auteurs ne distribuent pas ce lexique (même non achevé). Ils nous mettent l'eau à la bouche (BFQS ?) en nous laissant l'impression que leur énorme travail lexicologique de qualité est *béfracussé à la babounne*, expression figée de notre cru, N être béfracussé à la babounne, 'se dit d'un excellent travail dont personne ne profite des résultats'.

Michelle LECOLLE, Marie-Anne PAVEAU, Sandrine REBOUL-TOURÉ, Le nom propre en discours, *Presses Sorbonne nouvelle*, 2009, 216 pages, ISBN 978-2-87854-449-7.

Lu par **Maud EHRMANN**

European Commission - Joint Research Centre (JRC)

Cet ouvrage présente un ensemble de communications visant à rendre compte des approches discursives du nom propre (Npr). L'intention est ici de « donner la parole » à ce champ d'étude spécifique (et pour le moment peu exploré[2]) qu'est le Npr en discours en considérant un large éventail de Npr (non seulement les anthroponymes mais également les pseudonymes, toponymes, praxonymes et autres « polémonymes ») et en prenant appui sur deux approches complémentaires, la linguistique et l'analyse du discours.

Description générale

L'ouvrage s'articule en trois parties et comporte un avant-propos et une postface. La première partie, intitulée « Identité, identification et changement d'identité », propose une entrée en matière par le biais (classique) de la question du rapport entre le Npr et son référent. La seconde partie, rassemblant des études prenant appui sur des corpus de presse (« La construction de l'événement dans la presse : dénomination propre, nom propre d'événement »), s'intéresse, quant à elle, à l'analyse des divers éléments de nature discursive contribuant à la construction et/ou

2. Les auteurs de l'avant-propos qualifient même les approches discursives du nom propre de « balbutiantes » (p. 9).

modification du sens des Npr. La dernière partie, « Histoire, mémoire, légende », met l'accent sur une dimension peu considérée dans les études linguistiques sur le nom propre : la dimension « historico-discursive », ou comment le Npr peut devenir un lieu de mémoire pour telle ou telle communauté discursive.

Avant de rentrer dans le détail des contributions, il nous semble important d'évoquer le point suivant. Comme précisé dans le résumé, l'ouvrage aborde la problématique du Npr en discours selon deux approches, rapidement présentées dans l'avant-propos en tant qu'« analyse linguistique » d'une part, et « approche discursive », d'autre part. Pour fructueuse que soit la considération de ces approches au regard du thème abordé, leur absence de définition donne lieu, néanmoins, à quelque confusion. En effet, c'est au fur et à mesure que le lecteur comprend que le Npr est abordé tantôt par un spécialiste de l'analyse du discours, tantôt par un linguiste. La compréhension aurait sans doute été facilitée par un cadrage plus formel des perspectives d'étude dès le début de l'ouvrage. Dans sa postface, M.-N. Gary-Prieur souligne d'ailleurs d'amblée l'ambiguïté du mot *discours* présent dans le titre, ainsi que celle de l'expression *approches discursives*, et en détaille aussitôt les deux acceptions possibles : une *approche linguistique* considérant comme objet d'étude « le Npr en tant que forme de la langue » et une *approche discursive* considérant comme objet d'étude « un (des) discours sur le référent du Npr ». On peut donc regretter, à titre général, un manque de définition précise des champs d'action de chacune de ces disciplines.

Résumé des contributions

La première partie comporte deux contributions s'intéressant, chacune à sa manière, à la question de l'identité et de l'identification du référent désigné par le Npr. En effet, si l'article de G. Achard-Bayle pose la question suivante : si l'individu change d'identité, qu'advient-il de son nom ?, celui de G. Cislaru demande plutôt : si l'individu change de nom ou en prend un autre, qu'advient-il de son identité (ou : que nous dit-il de son identité) ? Revenons successivement sur ces études. G. Achard-Beyle analyse, au travers de l'étude de *cas* (au sens des *puzzling cases*) extraits de textes littéraires ou journalistiques, les diverses variations ou « degrés de résistance du Npr » face aux changements affectant son référent. En effet, face aux métamorphoses que peut subir le porteur du nom propre, ce dernier peut, dans les cas étudiés, soit se maintenir, soit au contraire se vider de son sens. Au final, s'il est possible d'observer que le Npr peut être rigide tout comme « particulièrement plastique », l'auteur reconnaît qu'il est encore difficile, à ce stade de son étude, de dégager des principes régissant ces phénomènes. Dans sa contribution, G. Cislaru s'intéresse quant à elle au Npr comme autonomination, avec une étude de l'apparition et de l'usage des pseudonymes dans les forums Internet. L'auteur, par la mise en valeur de leur triple niveau de signifiance (onomastique, personnel et communautaire) et de leur étroite dépendance vis-à-vis du discours (le pseudonyme existe pour un « qui écrit »), montre comment le sujet parlant, ou écrivant, peut se construire une identité (exclusivement) discursive.

Ces contributions, bien qu'abordant la question de manière opposée, montrent toutes deux que le rapport du nom propre à son référent est loin d'être univoque et que divers paramètres, notamment discursifs (au sens linguistique du terme), entrent en jeu dans sa détermination. Intéressante et de très bonne qualité, on peut néanmoins regretter que la contribution de G. Cislaru ne discute pas de manière plus approfondie la question du statut du pseudonyme vis-à-vis du nom propre. M.-N. Gary-Prieur revient d'ailleurs sur ce point dans la postface, indiquant que, selon elle, le pseudonyme ne peut être considéré comme une sous-catégorie du nom propre.

La seconde partie se concentre sur les usages du Npr dans la presse et étudie les constructions/modifications de son sens. M. Veniard étudie un type particulier de Npr : les dénominations propres de guerre. L'auteur, après avoir clarifié le statut linguistique des dénominations propres, analyse comment ces dernières sont « susceptibles d'investissements sémantiques en discours », en examinant successivement les aspects énonciatifs, discursifs et interdiscursifs de leurs emplois. Il ressort de cette étude minutieuse que le sens des dénominations propres de guerre, au-delà de leur aspect descriptif, se définit *via* leur actualisation discursive et interdiscursive et selon un point de vue énonciatif particulier. A. Krieg-Planque s'intéresse, de manière plus générale, aux noms propres d'événement. L'auteur examine deux points notamment : du point de vue des pratiques journalistiques tout d'abord, elle explicite l'importance du nom propre d'événement, se trouvant au cœur du contenu transmis par le journaliste et polarisant trois besoins médiatiques de catégorisation, d'analogie et de prototypicité. Du point de vue des pratiques discursives ensuite, elle montre comment un nom propre d'événement doit sans cesse être réinterprété en contexte et comment, après de multiples réinvestissements sémantiques, il peut parfois devenir jusqu'à inintelligible. M. Lecolle étudie pour sa part le nom propre *Outreau* dont elle examine le changement de sens en discours. Se plaçant dans un cadre théorique faisant la distinction entre *signification* (d'ordre structural) et *sens* (inscription en discours) du Npr, l'auteur montre le caractère composite du sens attaché au toponyme *Outreau* et se propose de l'analyser. Pour ce faire, elle détermine un certain nombre d'éléments (d'ordre sémantique, distributionnel, énonciatif, etc.) « potentiellement responsables » de l'évolution du sens de *Outreau* et les étudie dans un corpus de presse organisé chronologiquement. Cette étude, sur ce qui pourrait désormais devenir un cas d'école dans la littérature sur le Npr, constitue une parfaite illustration d'un « dialogisme de la nomination ».

Ces trois études, explorant chacune à leur manière le Npr d'événement, cherchent à rendre compte de l'épaisseur sémantique dont peut se charger un Npr. Par l'étude de divers mécanismes syntagmatiques, énonciatifs, discursifs et interdiscursifs, ces contributions montrent comment le sens d'un Npr peut être déterminé, modifié et remotivé au fur et à mesure de ses actualisations en discours, révélant par là même les divers remodelages du référent. Ces trois études constituent une excellente porte d'entrée pour l'analyse du Npr d'événement en particulier, et du Npr en discours en général.

La dernière partie de ce recueil aborde le Npr comme lieu de mémoire : parce que référant à tel type d'entité, utilisé de telle sorte en discours par telle communauté, un nom propre peut se charger d'une signifiance historique ou légendaire. I. Khmelevskaia analyse l'usage des anthroponymes par les commentateurs sportifs et montre comment, par la comparaison métaphorique, les jeux de mots et la pratique de l'intertexte, se construit « l'espace légendaire du jeu ». M. Kasper rend compte des références à *Marx* dans les discours de la presse estonienne aujourd'hui. Enfin, M.-A. Paveau introduit le polémonyme ou nom de bataille. Dans le cadre d'une analyse cognitive du discours, elle met en avant la nécessaire prise en compte des subjectivités individuelle et collective pour analyser le sens des Npr de bataille (et pas seulement), lesquels peuvent être comparés à des « badges » à fonction identificatrice pour tel ou tel groupe.

Cet ouvrage donne un bon aperçu des actuelles approches discursives du nom propre, considérant une approche tantôt 'linguistique', tantôt 'analyse du discours'. Comme signalé plus haut, on peut regretter le manque de précisions au regard de la définition de ces différentes approches qui, si elles se dégagent au fur et à mesure de la lecture du recueil, n'apparaissent qu'en postface. Hormis ce point, cet ouvrage constitue une lecture intéressante pour toute personne ayant déjà de bonnes connaissances sur le nom propre (ces contributions très spécialisées ne reviennent que rarement sur les « fondamentaux » et ne sont pas à conseiller comme première lecture sur le Npr). D'un point de vue plus pratique, l'ouvrage est facile à consulter, avec un index des notions et des résumés en français et anglais.

Sandra KÜBLER, Ryan McDONALD, Joakim NIVRE, *Dependency Parsing*, *Morgan & Claypool Publishers*, 2009, 114 pages, ISBN 9781598295962.

Lu par **Marie CANDITO**

Équipe-projet Alpage, université Paris 7

L'ouvrage est un excellent panorama des techniques d'analyse syntaxique en dépendances. Il présente et compare les différentes méthodes connues, les formats de données et les protocoles d'évaluation utilisés. Sont présentées séparément deux types de méthodes qui, bien que compatibles, sont le plus souvent mutuellement exclusives : celles utilisant une grammaire formelle de dépendances qui définit un ensemble d'énoncés couverts, et celles utilisant un apprentissage à partir de données linguistiques. Les auteurs se limitent à l'apprentissage supervisé, i.e. *utilisant des données syntaxiquement annotées. Le cœur de l'ouvrage est la présentation et la comparaison de deux types d'algorithmes pour l'apprentissage supervisé d'analyseur syntaxique en dépendances : les systèmes à transition et les systèmes fondés sur des graphes, dont Joakim Nivre et Ryan McDonald sont respectivement les acteurs majeurs.*

Résumé de l'ouvrage

L'analyse syntaxique en dépendances (*dependency parsing*) a gagné une grande popularité ces dix dernières années pour trois raisons principales, selon les auteurs : (1) les grammaires de dépendances semblent mieux adaptées pour la représentation de langues à ordre libre ; (2) les représentations en dépendances fournissent plus directement les structures argumentales et (3), en TAL, l'analyse syntaxique en dépendances, en particulier statistique, a prouvé son efficacité pour une grande variété de langues.

L'introduction présente très succinctement le concept linguistique de dépendances. Le chapitre 2 définit formellement la tâche d'analyse syntaxique en dépendances, comme la production à partir d'une chaîne segmentée d'un arbre de dépendances. Les propriétés formelles de ces arbres sont interprétées en termes linguistiques et computationnels.

Le chapitre 3 présente les systèmes à transitions (initialement proposés en 2003 pour l'anglais par Yamada et Matumoto, et une variante en 2003 pour le suédois par Joakim Nivre). Les auteurs définissent la dérivation d'un arbre de dépendances comme une succession de transitions entre des configurations d'une machine abstraite, en partant d'une configuration initiale, et en aboutissant à une des configurations finales autorisées. Un exemple typique parmi d'autres d'un tel système à transitions analyse une configuration comme un triplet [pile, *buffer*, ensemble de dépendances dérivées jusque-là], et distingue trois types de transitions possibles : lire le *buffer* (*SHIFT* de la tête du *buffer* vers la pile), ou bien ajouter une dépendance typée entre le mot en haut de la pile et le mot débutant le *buffer*, dans les deux sens possibles.

Étant donné ce modèle, l'algorithme d'analyse syntaxique déterministe est très simple : il débute en configuration initiale, applique la transition la plus appropriée possible à chaque configuration, s'arrête à la première configuration terminale rencontrée, et retourne l'arbre de dépendances, nécessairement projectif, constitué de l'ensemble des dépendances présentes dans cette configuration terminale. Un tel algorithme analyse la phrase *en un temps linéaire* (O(n) avec n la taille de la phrase). Toute la complexité du modèle est reportée dans l'apprentissage de paramètres permettant de choisir le type de transition à appliquer pour passer à la configuration suivante, étant donné la configuration courante. Pour cela, les arbres de dépendances du corpus d'apprentissage sont convertis en séquences de configuration, qui sont ensuite représentées par des traits plus généraux tels que la catégorie/le lemme/la forme du n-ième mot du *buffer* ou de la pile, la distance entre un mot en haut de la pile et un mot en tête de *buffer*, le nombre et le type de dépendants déjà dérivés pour un de ces mots, etc. Les données d'apprentissage sont alors prêtes pour entraîner un classifieur qui, étant donné les traits représentant une configuration, détermine le

type de transitions à appliquer pour avancer dans l'analyse syntaxique[3]. Les auteurs citent, plus qu'ils ne détaillent, deux classifieurs qui ont typiquement été utilisés pour cette tâche :

(i) un classifieur avec apprentissage à base de cas (*memory-based learning*) et une classification de type k-plus proches voisins. La phase de classification est computationnellement plus coûteuse que l'apprentissage ;

(ii) les machines à vecteur de support (SVM), typiquement avec noyaux polynomiaux d'ordre 2, qui implique l'utilisation implicite systématique de paires de traits. Ils donnent les meilleurs résultats, mais, à l'inverse, l'apprentissage est très coûteux et nécessite souvent de séparer les configurations à classifier en x paquets et entraîner x classifieurs différents.

Sont également mentionnées des variantes pour les types de transitions possibles et pour l'algorithme d'analyse syntaxique, et des extensions pour traiter la non-projectivité.

Le chapitre 4 présente des méthodes dites « à base de graphes », car elles réutilisent des algorithmes bien connus en théorie des graphes pour le problème de l'analyse syntaxique en dépendances. La notion centrale est de formuler le score d'un arbre de dépendances comme une fonction (en général la somme) des scores de certains de ces sous-graphes. Le cas le plus simple est de prendre la somme des scores de chaque arc. Le problème de l'analyse syntaxique se ramène à trouver l'arbre maximisant ce score, ce qui est connu en théorie des graphes comme la recherche de l'arbre couvrant de poids maximal (*Maximal Spanning Tree, MST*) : on part du graphe connectant tous les mots dans tous les sens avec toutes les relations possibles. L'arbre de score maximal, éventuellement non projectif peut être obtenu avec l'algorithme Chu-Liu-Edmonds. Une variante par Eisner de l'algorithme CKY, bien connu en analyse syntaxique en constituants, fournit l'arbre projectif de score maximal. L'apprentissage dans ce cadre consiste à apprendre les scores à affecter aux arcs. Les arcs sont transformés en vecteurs de traits, incluant typiquement le label de la dépendance, la catégorie des mots précédents/suivants le gouverneur/le dépendant, la distance entre ces mots, etc. L'apprentissage du vecteur de poids peut être fait, par exemple, avec un algorithme inférentiel de type « perceptron ».

Cependant, considérer les arcs isolément n'est pas linguistiquement satisfaisant. Un modèle étendu utilise comme score la somme des scores des paires d'arcs adjacents, au lieu des simples arcs. Maximiser ce score devient malheureusement NP-complet dans le cas non projectif, mais reste polynomial dans le cas projectif[4].

---

3. Ainsi, si formellement dans ce modèle le choix d'une transition ne dépend que de la configuration courante, les choix précédents sont reflétés dans les dépendances déjà dérivées.

4. Un article cité donne une complexité de O(n3) pour ce cas, ce que le livre n'indique pas.

Les méthodes avec grammaire sont présentées au chapitre 5. Une première approche montre une grammaire formelle en dépendances, convertie en grammaire hors contexte, sur laquelle les algorithmes classiques d'analyse syntaxique hors contexte peuvent être appliqués. Dans une seconde approche, l'analyse syntaxique est un problème de satisfaction de contraintes, celles-ci formant la « grammaire ». Une analyse syntaxique par propagation de contraintes pondérées est possible mais ne permet pas toujours une désambiguïsation totale. Une technique par réparation part d'un arbre initial arbitraire, corrigé pas à pas d'après les poids des contraintes.

Le chapitre 6 détaille les mesures d'évaluation des analyseurs syntaxiques en dépendances, ainsi qu'une procédure classique de conversion d'arbres de constituants en arbres de dépendances projectifs, typés ou non selon les informations de fonctions grammaticales disponibles dans les arbres de constituants de départ. Ce chapitre contient également une description de la campagne d'évaluation de l'analyse syntaxique multilingue en dépendance proposée en 2006 et 2007 à la conférence CoNLL, où les participants étaient invités à entraîner des analyseurs syntaxiques sur des corpus pour différentes langues, représentant, en 2007, neuf familles linguistiques. L'ouvrage fournit une analyse syntaxique intéressante des résultats d'après la typologie des langues à traiter.

Le chapitre 7 présente une comparaison des différentes méthodes. Celles par transitions permettent des traits assez riches, mais souffrent de la propagation d'erreurs faites lors des premières transitions. À l'inverse les méthodes par graphes fournissent une solution optimisée globalement, mais au prix de traits assez pauvres. La performance moyenne sur plusieurs langues est similaire, mais on trouve des différences pour une langue donnée. Pour les méthodes par grammaire formelle ou par contraintes, l'ouvrage établit des liens formels avec les méthodes par apprentissage.

Commentaire

L'ouvrage est très complet et à la fois clair et concis, avec une volonté de capturer sous un cadre unifié les différentes techniques présentées, ce qui fournit un recul par rapport aux détails de chaque méthode. Ma seule critique serait peut-être un trop grand « œcuménisme » qui rend difficile le fait de repérer les forces et faiblesses de chacune des méthodes. Aucune information de performance n'est donnée pour les méthodes symboliques, et aucune comparaison de temps de traitement n'est fournie, ce qui serait pourtant intéressant (*cf.* les complexités très différentes des algorithmes présentés). Enfin, on ne trouvera pas non plus de comparaison de performance avec les analyseurs syntaxiques en constituants.

Kam-Fai WONG, Wenji LI, Ruifeng XU, Zheng-sheng ZHANG, *Introduction to Chinese Natural Language Processing*, *Morgan & Claypool publishers*, 2010, 148 pages, ISBN 9781598299328.

Lu par **Liangcai SHEN**

Doctorant, SYLED/CLA2T. Université de la Sorbonne nouvelle Paris 3

Comme en attestent les nombreuses critiques favorables qui ont accompagné la publication de l'ouvrage, les communautés scientifiques occidentales attendaient depuis longtemps un texte présentant les nouvelles méthodes et normes d'excellence en ingénierie multilingue chinoise. Les auteurs ont réussi à articuler des savoirs statistiques, algorithmiques et des connaissances sur les bases linguistiques du TAL dans un livre équilibré, complet et facile à lire.

Structure de l'ouvrage

L'ouvrage est centré sur l'analyse morphologique du chinois, qui est à la base du TAL dans le domaine sinophone. Après un chapitre introductif, il est divisé en trois parties. Le chapitre 2 présente les concepts de *caractères*, *morphèmes* et *lexèmes* chinois, sous un angle linguistique. Le chapitre 3 décrit les caractéristiques des mots chinois en liaison avec les applications du TAL. Le chapitre 4 aborde les solutions techniques pour la segmentation en *mots*. Il est suivi par un exposé consacré à la détection de mots inconnus (chapitre 5). Le chapitre 6 introduit la notion de *sens* des mots et présente plusieurs ressources linguistiques consacrées au chinois. Enfin, les chapitres 7 et 8 abordent les questions de sémantique fondées sur l'extraction de *collocations*.

Contenu de l'ouvrage

Le chapitre 2 décrit les unités morphologiques : mots et morphèmes du chinois, depuis les processus de formation des mots jusqu'à leur découpage automatique dans la chaîne écrite.

Le chapitre 3 est consacré aux caractéristiques linguistiques et textuelles du chinois susceptibles de poser des défis particuliers lors des traitements automatiques : les variations régionales et/ou stylistiques qui peuvent exister parmi les caractères et leur encodage, les conventions textuelles spéciales de l'imprimerie et de la ponctuation, les nombreuses ambiguïtés et l'absence de marquages grammaticaux clairement définis pour de nombreux mots inconnus, comme les noms, les abréviations et les translittérations.

Le chapitre 4 aborde les problèmes de la segmentation du mot chinois. L'auteur commence par noter les différences entre le chinois et l'anglais en donnant une définition formelle de la segmentation du mot chinois. Puis, quelques exemples concrets sont introduits afin de mettre en évidence les deux principaux défis du

TAL, à savoir : le traitement des ambiguïtés et des mots inconnus. Les différents algorithmes spécifiques de segmentation sont ensuite soumis à une classification. Trois normes d'évaluation pour la segmentation sont ensuite décrites ainsi que les résultats des évaluations de la segmentation lors du concours SIGHAN[5]. Les résultats sont évalués selon cinq critères : rappel, précision, F-mesure, rappel des mots inconnus et rappel des mots connus. La fin du chapitre est consacrée à la présentation de deux outils libres pour la segmentation du mot chinois.

Le chapitre 5 est consacré aux mots inconnus qui posent des problèmes importants lors de la segmentation en mots des textes chinois. Les noms propres qui réfèrent à des personnes, des lieux, des noms d'organisation constituent des sources importantes de mots inconnus. Les méthodes de reconnaissance des noms utilisent des indices tels que la structure commune de nommage, le repérage de caractères utilisés régulièrement dans les noms, la prise en compte des données contextuelles, etc. Ces indices sont obtenus par extraction manuelle ou automatique à partir de corpus de textes.

Le chapitre 6 décrit trois ressources linguistiques chinoises importantes. Ces ressources sont précieuses pour de nombreuses applications du TAL du chinois, telles que la recherche d'information fondée sur les concepts, l'extraction, la classification de documents, la désambiguïsation, et la traduction automatique. D'autres ressources similaires sont exposées dans l'annexe A : dictionnaires imprimés, lexiques électroniques, corpus, transcriptions, etc. Certaines de ces ressources sont librement accessibles sur Internet.

Le chapitre 7 présente les concepts de base qui ont trait au repérage des collocations en chinois. Le terme « collocation » ayant différentes définitions qui font débat dans la littérature. Ces différentes définitions de la collocation sont examinées au début du chapitre. Elles sont suivies par un recensement des caractéristiques qualitatives et quantitatives des collocations. Les schémas de catégorisation de collocations sont ensuite passés en revue ainsi que les apports résultant de la prise en charge des collocations dans les applications pratiques.

Le chapitre 8 examine les principales approches permettant l'extraction de collocations dans les textes chinois. Ces approches s'appuient sur différentes propriétés des collocations afin de mieux les repérer. Parmi elles, l'approche statistique, dite *fenêtre de contexte*, qui est à la base de la plupart des systèmes d'extraction des collocations. À ce propos, la fréquence des cooccurrences des collocations reflète leur propriété récurrente alors que celle de leur distribution laisse observer des limites dans leur variation compositionnelle. L'approche fondée sur la syntaxe souligne que les collocations bigrammes doivent être syntaxiquement

5. *Special Interest Group on Chinese Language Processing* : le 2e concours SIGHAN a eu lieu à l'occasion de la conférence de l'*Association for Computational Linguistics* (ACL) en 2003. http://sighan.cs.uchicago.edu/constitution.htm.

dépendantes, ce qui permet également d'affiner l'extraction de collocations. L'approche sémantique est élaborée sur la base des observations selon lesquelles les composants d'une collocation peuvent être remplacés par un ensemble restreint de mots similaires ou équivalents. Une approche hybride fondée sur la catégorisation est alors étudiée. Elle montre son efficacité et ouvre de nouveau champs pour de futures études dans l'extraction des collocations. Enfin, deux sources de référence disponibles sont introduites pour l'évaluation de l'extraction automatique de collocations.

Conclusion

Malgré quelques erreurs dans les références bibliographiques, ce livre, conçu dans un esprit synthétique, fournit un état de l'art du domaine dans toute sa richesse tant du point de vue des connaissances rassemblées que des méthodologies. Il comprend également des exemples issus de certains dialectes chinois, et pourrait constituer un outil indispensable pour les étudiants souhaitant effectuer une recherche scientifique.

Ce livre est le premier à répondre aux besoins du public sinophone. Les lecteurs familiarisés avec la langue chinoise y trouveront une présentation accessible à la modélisation, les méthodes, les algorithmes et les techniques. Il permettra à ceux qui sont dotés d'une bonne formation en linguistique de mieux comprendre les bases techniques de l'ingénierie linguistique. À l'inverse, les initiés en traitement informatique peuvent en apprendre davantage sur la syntaxe, la sémantique et l'analyse du discours du chinois. Au-delà des cercles de recherches universitaires, les chercheurs, les développeurs et les techniciens du TAL appliqué au chinois, ainsi que ceux qui travaillent dans des disciplines connexes, devraient trouver dans ce livre une aide précieuse pour se familiariser avec le domaine en plein essor des technologies langagières.

Laurence ROSIER, Le discours rapporté en français, *Ophrys*, 2008, 148 pages, ISBN 978-2-7080-1214-1.

Lu par **Denis LE PESANT**

MoDyCo (Université Paris Ouest Nanterre)

Le discours rapporté en français, de Laurence Rosier, constitue une synthèse claire et fidèle des études les plus récentes des divers types de discours rapportés. De ce fait, son ouvrage intéressera les étudiants, enseignants et chercheurs en analyse littéraire et en linguistique du français.

L'auteur ne se contente pas d'évoquer les travaux des autres (notamment Jacqueline Authier, mais aussi Oswald Ducrot, l'École de Genève, la SCAPOLINE et Marc Wilmet parmi bien d'autres). Elle apporte aussi sa propre contribution à

l'analyse du phénomène. Elle propose un système des discours rapportés organisé autour de la « fracture linguistique » que constitue l'opposition entre le discours indirect et le discours direct. Elle schématise le système sous la forme d'un V dont la pointe figure ladite césure.

Sur la branche gauche du V, qui est l'axe du discours indirect, on s'élève du moins libre au plus libre et au plus dialogique : discours indirect sans *que*, discours indirect « mimétique » (ex. : *Il disait « que sa maladie le hantait »*), discours indirect avec incise, discours indirect avec *que*, discours indirect libre mimétique (ex. : *Il n'arrêtait pas de parler : nom de nom, comme sa maladie le hantait !*), relative de discours indirect libre (ex. : *Il n'arrêtait pas de parler de sa maladie, qui le hantait*). Symétriquement, sur la branche droite du V qui est l'axe du discours direct, on part du discours direct avec *que* (ex. : *Il n'arrêtait pas de dire que « ma maladie me hante »*), on monte vers le discours direct prototypique (ex. : *Il n'arrêtait pas de parler : « ma maladie me hante »*), puis vers le discours direct émancipé typographiquement (ex. : *Il n'arrêtait pas de parler ma maladie me hante*), pour aboutir au discours direct libre (ex. : *Il la regarda. Ma maladie me hante*).

Mais il ne s'agit là que de onze jalons d'une série que Laurence Rosier conçoit comme *continue*. Le modèle de l'auteur combine donc la représentation d'une dichotomie (discours indirect *vs* discours direct) avec la représentation d'un continuum. Cette solution lui permet de capturer dans le même modèle la multitude des cas de figure offerts par les corpus (littérature, presse, textes scientifiques, oral, blogs, etc.) et marqués par l'hybridation des types.

Le dernier chapitre est particulièrement heureux. Il met l'accent sur la nécessité de rapporter la diversité des phénomènes à celle des genres de discours. Y sont évoqués notamment la citation dans les discours scientifiques, le discours rapporté dans l'argumentation, la traduction comme discours rapporté, les formes récursives dans le récit et l'autocitation.

La sensibilité aux aspects polyphoniques et dialogiques des discours ne conduit pas Laurence Rosier à sous-estimer les aspects syntaxiques. La prise en compte de l'oral et celle de la diversité des genres de discours tranchent avec une tradition qui a eu le tort, sur la question du discours rapporté, de privilégier l'écrit et plus particulièrement l'écrit littéraire[6]

6. La révision matérielle de l'ouvrage laisse à désirer. Par exemple, il n'y a pas moins de neuf « coquilles » dans les huit pages de l'introduction. Il est à souhaiter qu'une réédition soit l'occasion de corriger ce défaut.

Graham WILCOCK, *Introduction to Linguistic Annotation and Text Analytics*, *Morgan & Claypool publishers*, 2009, 159 pages, ISBN 9781598297386.

Lu par **Lydia-Mai HO-DAC**

Université Catholique de Louvain-la-Neuve (UCL)

Cet ouvrage didactique vise à guider l'étudiant ou le chercheur (ou encore le « commercial » selon les dires de l'auteur) dans ses premiers pas en annotation linguistique assistée. L'organisation de l'ouvrage est progressive : tout d'abord une introduction au langage XML pour la constitution du corpus et des ressources, suivie d'une présentation d'outils d'annotation (WordFreak) et de ressources pour l'annotation manuelle (Penn Treebank) ou assistée (OpenNLP), pour finir sur une présentation de plates-formes TAL (GATE et UIMA) donnant la possibilité d'intégrer les annotations semi-manuelles à une chaîne de traitements plus complexes permettant entre autres des techniques d'apprentissage et d'extraction d'informations. Chaque étape de cette progression est illustrée par des mises en pratique immédiates : de l'installation des logiciels – quasiment tous libres – à l'utilisation de systèmes d'annotation automatique et de traitements spécifiques (tokeniser, étiquetage morphosyntaxique, reconnaissance d'entités nommées, résolution d'anaphores, etc.).

Chapitre 1. Introduction à XML

Dans ce chapitre, l'auteur présente les bases du langage XML et XSLT pour la construction du corpus d'étude. Cette introduction, très pratique, pour qui n'a jamais manipulé XML et XSLT, se fait *via* l'utilisation de l'éditeur jEdit et s'achève sur une valorisation de l'annotation *stand-off* et un très bref aperçu de la plate-forme GATE pour gérer des annotations au format XML.

Chapitre 2 et 3. Quelles annotations en linguistiques ?

L'auteur dresse ici un panorama d'exemples d'annotations à différents niveaux d'analyses : de la simple segmentation (en tokens et phrases) vers des annotations de plus en plus complexes, catégories morphosyntaxiques, analyse syntaxique, distinction prédicats/arguments, connecteurs (selon le modèle d'annotation du Penn *Discourse Treebank*), relations de coréférence et entités nommées. L'auteur utilise l'interface WordFreak pour appliquer ces niveaux d'annotations à un exemple commun (le sonnet 130 de Shakespeare). En début de chapitre, seules des annotations purement manuelles sont présentées. Elles sont ensuite reprises, en deuxième partie de chapitre et dans le chapitre 3, en version « préannotation automatique à valider manuellement » (les traitements automatiques utilisés proviennent des projets OpenNLP et Standford NLP).

Chapitre 4. Transportabilité des annotations

Ce chapitre est consacré aux fichiers XSLT (introduits très rapidement dans le chapitre 1) et fournit de nombreux exemples de transformations XSLT, notamment

pour passer des sorties/entrées des traitements OpenNLP (basés sur textes bruts) aux formats XML requis par GATE ou WordFreak, ou encore pour échanger les annotations entre GATE (format XML GATE détaillé) et WordFreak afin, par exemple, d'éditer les annotations issues de GATE *via* WordFreak. Ce chapitre se finit avec la présentation et la promotion d'un format standard pour l'échange de données XML : XMI (XML Metadata Interchange).

Chapitre 5. GATE et IUMA, plates-formes pour le TAL

Ce chapitre compare deux plates-formes dédiées aux TAL et à l'annotation de corpus : GATE et UIMA (qui utilise le format XMI). Bien que GATE ait déjà été présenté dans les précédents chapitres, l'auteur fait ici une introduction détaillée pas à pas de cette plate-forme et de deux outils associés : le module d'extraction d'informations ANNIE, quelques règles d'extraction de patrons en JAPE et les lexiques disponibles ou à construire). La même organisation est suivie pour présenter la plate-forme UIMA, ses *plugins* et ses dictionnaires/lexiques.

Chapitre 6. Extraction d'informations

Ce dernier chapitre propose quelques pistes d'outils, pas toujours libres et gratuits, pour l'annotation d'éléments issus d'une analyse textuelle[7] (« *text analytics* »). Après avoir brièvement expliqué les différences entre techniques d'apprentissage et application d'un lexique, l'auteur nous liste les différents outils permettant la reconnaissance d'entités nommées, la résolution de coréférences et l'extraction d'informations. Les exemples de ce chapitre, qui portent alors sur un texte long non poétique, sont beaucoup moins commentés que dans les autres chapitres et donnent rarement l'occasion à des travaux pratiques. Il s'agit davantage d'un panorama de ce qui existe et est reconnu actuellement.

Conclusion

Cet ouvrage, d'une grande qualité pédagogique, permet de sauter à pieds joints dans le monde de l'annotation de corpus, fournissant une bonne palette d'outils et de conseils au novice en annotation de corpus. Les nombreuses copies d'écrans aident considérablement la prise en main de cet ouvrage. Cet aspect très illustratif entrave grandement toute lecture « théorique » (sans « mettre les mains à la pâte »). Le grand nombre de figures, le fait qu'elles se trouvent souvent sur une autre page que celle concernée et leur délimitation mal marquée visuellement, entraînent une lecture « linéaire » peu agréable. Il faut lire ce livre en « faisant ».

Très bien référencé (chaque étape a son lot d'ouvrages dédiés et de liens vers les logiciels à télécharger librement), l'auteur offre la possibilité d'accéder facilement à une littérature plus détaillée sur le langage XML et XSLT, les formats d'annotation, les outils d'annotation, etc. Les références données permettent surtout d'envoyer le

7. Il ne s'agit pas d'une analyse linguistique mais de traitements automatiques visant la reconnaissance de phénomènes d'ordre discursif (*i.e.* au-delà de la syntaxe).

lecteur vers d'autres modèles d'annotation et d'analyse textuelle que ceux, relativement simples linguistiquement parlant, proposés dans cet ouvrage davantage didacticiel qu'informatif.

En conclusion, cet ouvrage constitue un très bon manuel à utiliser en support de cours ou en guide pour se lancer dans un travail d'annotation « assistée », et j'insiste sur ce terme, car l'annotation est ici principalement expliquée dans sa partie logiciel et non linguistique (linguistique descriptive ou linguistiques de corpus) : pas d'exemples d'annotation de corpus (toujours un seul texte), aucune remarque sur la masse de données que peut représenter une campagne d'annotation portant sur un vrai corpus. L'annotation est toujours envisagée dans un but applicatif plutôt que exploratoire.