

---

## Notes de lecture

### Rubrique préparée par Denis Maurel

Université François Rabelais Tours, LI (Laboratoire d'informatique)  
denis.maurel@univ-tours.fr

---

**Madjid IHADJADENE, Méthodes avancées pour les systèmes de recherche d'informations, Hermès Lavoisier, 2004, 246 pages, ISBN 2-7462-0846-6.**

par Nicolas HERNANDEZ

GREYC - DoDoLa  
nicolas.hernandez@info.unicaen.fr

---

*Ce traité s'adresse aux chercheurs et aux ingénieurs désireux de s'approprier des technologies et des méthodes avancées pour le traitement de problématiques liées à « l'accès à l'information ». Outre la communauté en Recherche d'Information (RI), les communautés en Traitement Automatique des Langues (TAL), en Interface Homme-Machine (IHM), en Ingénierie Documentaire, en Web Sémantique et en Base de Données (BD) sont représentées et concernées.*

*L'ouvrage traite de sujets qui dépassent le cadre traditionnel de la RI (c'est-à-dire selon une approche par interrogation qui donne un accès formel et analytique à un fond documentaire à partir de la formulation de requêtes descriptives et dont la satisfaction de l'utilisateur est mesurée en terme de précision et de rappel). Différents aspects d'une activité de recherche sont développés à la lumière de caractéristiques (cognitives) des utilisateurs, de propriétés (linguistiques et structurelles) de données et du type de recherche d'information considéré (question-réponse, RI par navigation, ...).*

*Dix-sept auteurs francophones se répartissent sur 10 chapitres de 24 pages en moyenne. Chaque chapitre donne les clefs d'un champ de recherche en présentant ses enjeux et ses problématiques. Leur corps est constitué de l'exposé des techniques majeures du domaine. Des illustrations et des exemples nombreux et pertinents viennent enrichir les descriptions et les explications. Les dernières sections traitent généralement de questions d'évaluation et de prospective. Les chapitres se terminent avec une bibliographie très complète du domaine.*

De par la diversité des recherches présentées, une organisation globale a du mal à se dégager. Néanmoins chaque chapitre peut se lire de manière indépendante. Une table des matières détaillée et un index général permettent des lectures ciblées dans l'ouvrage. Nous donnons ici un bref aperçu du contenu de chaque chapitre.

Dans le premier chapitre, Imad Saleh et Fabrice Papy se focalisent sur la recherche d'information en terme de navigation (exploration libre et informelle) comme moyen d'accès aux informations d'un hypertexte. Re-situant leur présentation par rapport aux caractéristiques structurelles des hypertextes et aux capacités cognitives des utilisateurs, ils énoncent des principes de conception d'outils de parcours et de visualisation de données. Une dernière partie concerne les techniques d'aide au choix de liens et au calcul de recommandation.

L'essentiel du second chapitre, rédigé par Jacques Le Maître, Elisabeth Murisasco et Emmanuel Bruno, présente de manière concise et détaillée les langages XPath et XQuery dont les auteurs montrent, au travers d'exemples, l'intérêt pour la recherche d'information dans des documents décrits en XML. Le chapitre se conclue par un état de l'art commenté des logiciels de stockage et d'interrogation de données XML actuels, avec un rapprochement avec les BD relationnels.

Le troisième chapitre porte sur les langages de métadonnées et les différents Webs Sémantiques qu'ils sous-tendent. Après avoir introduit les motivations et l'infrastructure du Web Sémantique telles que proposées par le W3C, Helka Folch et Benoît Habert centrent leur présentation sur la description et la mise en parallèle des langages RDF et Topics Maps autant du point de vue technique que de leur expressivité ; le premier semblant plus désigné pour soutenir une description formelle et consensuelle des ressources tandis que le second étant davantage adapté pour soutenir la coexistence de points de vues sur de l'information. Ils terminent en parlant d'un troisième aspect du Web Sémantique : le niveau ontologique des métadonnées (à travers OWL).

Dans le quatrième chapitre, Patrice Bellot décrit d'une part les principes de méthodes de classification (partitionnement, hiérarchie, cartes auto-organisées) et de catégorisation (k plus proches voisins, bayésienne naïve, machines à vecteurs supports, arbres de décision, entropie maximale, ...) de données. D'autre part, il traite de techniques d'enrichissement de requêtes (expansion de requêtes) à partir de documents trouvés ou de connaissances externes.

Au chapitre cinq, Mountaz Hascoët discute de paradigmes de visualisation d'information permettant des « vues d'ensemble ». Plusieurs techniques sont présentées selon que les données soient multi-dimensionnelles, structurées en arbre, en graphe, ou décrites en termes de « similarités » avec d'autres données. Une seconde partie du chapitre traite de procédés d'interaction, c'est-à-dire de l'exploitation des vues d'ensemble (filtrage sémantique, multi-niveaux de détails, zoom infini et sémantique). Chaque procédé est discuté relativement à la structure et à la taille des données qu'il manipule, à sa facilité de mise en place, à son usage pour des tâches interactives, aux effets perceptifs qu'il produit, etc.

Le chapitre six, rédigé par Christian Fluhr, traite de la recherche d'information interlingue. Après avoir exposé des solutions aux problèmes linguistiques (codage, flexions, unité lexicale) des systèmes multilingues (système qui peut, avec des ressources linguistiques différentes, traiter alternativement des langues

correspondantes), l'auteur présente les différentes approches des systèmes d'interrogation interlingue. Sa contribution s'axe sur les mécanismes d'une approche à base de dictionnaires bilingues (sur le couple français-anglais), dont il décrit les procédés de production à partir de corpus parallèles ou comparables.

Le chapitre sept, co-écrit par Mohand Boughanem, Mohamed Tmar et Hamid Tebri, définit la tâche de filtrage d'information comme de la recherche d'information sur des collections dynamiques (arrivée en continu de nouveaux documents) à partir de profils thématiques connus et stables. La plus grande partie du chapitre est dédiée à l'exposé de modèles mathématiques quant à la résolution des problèmes majeurs du domaine : l'apprentissage des profils (similaires aux principes de reformulation de requêtes) et celui des seuils de décision de filtrage.

Dans le huitième chapitre, Jacques Savoy met en lumière des différences essentielles entre le Web et un fond documentaire traditionnel. L'auteur rappelle d'abord quelques caractéristiques statistiques du graphe que constitue le Web, puis il explique les principes et le fonctionnement des moteurs de recherche de première génération (s'appuyant uniquement sur le texte des pages afin de dépister l'information souhaitée) et de deuxième génération (utilisant les hyperliens afin d'accroître la qualité des réponses ; par exemple à l'aide d'activation propagée, de l'algorithme de Kleinberg ou celui du PageRank).

Le chapitre neuf, rédigé par Brigitte Grau, traite d'une recherche d'information où le besoin de l'utilisateur n'est satisfait qu'après récupération d'une réponse concise à une requête interprétée comme une question. L'auteur retrace l'évolution historique des systèmes et des enjeux du Question-Réponse (Q-R) notamment au sein des campagnes d'évaluation TREC. L'essentiel du chapitre décrit le fonctionnement d'un système de Q-R ainsi que les différents processus de traitement linguistique touchant aux différents niveaux d'analyse des questions, du traitement des documents et de l'extraction des réponses.

Dans le chapitre dix, Laurence Favier et Madjid Ihadjadene présentent une étude des solutions techniques du marché dans le domaine de la veille et d'intelligence économique (dont l'objectif général est d'éclairer la prise de décision en management). Différentes activités sont analysées : veille stratégique, intelligence économique, gestion de connaissances, techniques de fouilles de données et de découvertes de connaissances. Confrontant les pratiques de veille entre les entreprises françaises et américaines, les auteurs observent l'inadéquation entre l'offre technologique et le cycle de veille en constatant que tout en ouvrant sur de nouvelles performances, la technologie reste inadaptée aux besoins qu'elle contribue à développer.

En résumé, le traité présente deux particularités majeures : d'une part, de décrire des savoirs et des savoirs-faires avancés dans différents champs de la RI, et d'autre part de très bien les exposer. Plus qu'un accès à des connaissances, le traité nous offre les moyens de se les approprier ainsi que de les mettre en application. Par

ailleurs la réunion de ces divers sujets écrits par des chercheurs distincts, experts de leur domaine, nous faisant profiter de leur expérience, constitue un apport certain.

Ces caractéristiques assurent au traité sa place au sein des ouvrages en RI. Notons qu'il étend le contenu d'un premier tome (dirigé aussi par M. Ihadjadene), lequel traite des modèles d'indexation de l'information dans le cadre d'une RI traditionnelle. En comparaison avec les ouvrages de Baeza-Yates et Ribeiro-Neto, (1999) et de Chowdhury (2005) ayant pour objectif de couvrir les domaines de la RI des techniques de bases aux avancées, le traité présente l'intérêt de réactualiser bon nombres des sujets ainsi que d'en introduire de nouveaux (RI par navigation, recherche dans des documents XML, Web Sémantique, multilinguisme, question-réponse, outils de veille, filtrage d'information). Le manuel de Lallich-Boidin et Maret (2005) quant à lui s'axe sur la présentation des bases de la linguistique et de la logique informatique pour comprendre les traitements auxquels les requêtes et les textes sont soumis.

L'une des critiques que l'on pourrait formuler est que de nombreux thèmes de RI se trouvant traités sur plusieurs chapitres, l'ouvrage gagnerait à se doter d'un index thématique plus structuré afin de faciliter des lectures transversales du traité, sur des thèmes tels que « l'usage du TAL en RI », « la place de l'utilisateur dans le processus de recherche », ou encore « les techniques de formulation des besoins ».

Parmi les sujets que le traité n'aborde pas et que nous souhaiterions voir dans de futures éditions, nous comptons les questions d'évaluation orientées utilisateur, les techniques de RI intra-documentaire, en bases documentaires multimédia, et pour documents composites.

En conclusion, c'est un ouvrage que nous avons très apprécié et nous en recommandons la lecture.

---

**Geneviève LALLICH-BOIDIN, Dominique MARET, Recherche d'information et traitement de la langue : fondements linguistiques et applications, Presses de l'enssib, 2005, 288 pages, ISBN 2-910227-60-X.**

**par Dominic FOREST**

*Observatoire de Linguistique Sens-Texte (OLST), Université de Montréal  
dominic.forest@umontreal.ca*

---

*L'ouvrage de Geneviève Lallich-Boidin et de Dominique Maret est une introduction aux concepts principaux et aux techniques fondamentales du traitement automatique de la langue dans leur application à la recherche d'informations à partir de documents textuels.*

L'ouvrage est divisé en deux parties. La première, composée des chapitres 1 à 7, aborde les fondements théoriques de la recherche d'informations en lien avec le

traitement automatique du langage. Le premier chapitre traite de la segmentation. Cette opération préliminaire nécessaire à tout traitement informatisé des documents textuels consiste à identifier, au sein d'un corpus de documents, les différents éléments (il s'agit traditionnellement de mots ou de phrases) qui servent d'ancrage à la recherche d'informations. Ce chapitre identifie clairement les nombreux enjeux linguistiques soulevés par cette opération. Le deuxième chapitre poursuit la réflexion entamée dans le chapitre précédent en traitant de la question de la lemmatisation des données textuelles. Après avoir présenté comment peut être réalisée l'opération de segmentation d'un document, ce chapitre s'attarde à l'opération consistant à réduire et à organiser les différents mots extraits. Ce deuxième chapitre présente clairement les concepts principaux auxquels fait appel l'opération de lemmatisation, tout en s'attardant sur l'opération d'étiquetage morphosyntaxique nécessaire à la lemmatisation.

Le troisième chapitre est consacré à l'analyse syntaxique des syntagmes pouvant être extraits à partir de documents. L'objectif de ce chapitre est de démontrer comment il est possible d'extraire, grâce à certaines structures linguistiques, des termes candidats décrivant adéquatement le contenu des documents. Le quatrième chapitre est consacré aux enjeux sémantiques et pragmatiques dans le cadre de la recherche d'informations.

Le cinquième chapitre, étroitement relié à celui sur la lemmatisation, traite de la question de l'affixation. Cette dernière est abordée sous les angles morphologique, syntaxique et sémantique. Le traitement informatique de l'affixation est essentiel dans le domaine de la recherche d'informations, car il permet de reconnaître certains termes sémantiquement reliés partageant un morphème commun. Le sixième chapitre traite des enjeux terminologiques. Ce chapitre démontre bien l'importance que l'on doit accorder aux termes complexes en recherche d'informations. Les auteurs présentent de manière succincte trois principales méthodes permettant d'assister l'identification de termes complexes dans un corpus de documents (méthodes des cooccurrences, des n-grammes et des segments répétés).

Le chapitre 7, le plus théorique de l'ensemble de l'ouvrage, aborde la question des grammaires de réécriture. La présentation des concepts principaux des grammaires de réécriture est effectuée de manière juste, mais les auteurs font appel à des concepts et à des techniques avec lesquels les lecteurs auxquels s'adresse ce livre ne sont certainement pas familiers (automate à états finis, automate à pile, etc.). Par ailleurs, contrairement aux chapitres précédents, le lien entre les concepts abordés dans ce chapitre et la problématique de la recherche d'informations n'est pas des plus évidents.

La seconde partie de l'ouvrage, regroupant les chapitres 8 à 12, illustre à l'aide d'exemples détaillés les notions explicitées dans la première partie de l'ouvrage. Ainsi, le chapitre 8 s'attarde à la problématique complexe de l'identification des noms propres. À cet égard, les auteurs identifient clairement les défis à relever

(erreurs orthographiques, problèmes découlant de la correspondance entre l'oral et l'écrit, etc.) lors de l'implémentation informatique de cette opération.

Dans le chapitre suivant, les auteurs présentent les principales opérations lors de la recherche d'informations en langue naturelle. Dans un premier temps, les principaux concepts de la recherche en texte intégral (indexation, recherche booléenne, etc.) sont présentés et illustrés par des exemples simples. Dans un deuxième temps, les auteurs présentent une démarche d'analyse d'une requête en langue naturelle. Cette seconde section du chapitre 9 constitue un très bon exemple d'application des concepts et des techniques exposés précédemment. On y retrouve une démonstration de la pertinence réelle des opérations de segmentation, de lemmatisation et d'analyse morphosyntaxique dans le cadre d'un processus de recherche d'informations textuelles. Finalement, la troisième partie de ce chapitre est consacrée à la traduction d'une requête en langue naturelle adressée à un système de recherche en texte intégral.

Le chapitre 10, intitulé « Ressources linguistiques multilingues », vise d'abord à présenter l'architecture d'un dictionnaire adaptée à la recherche d'informations translinguistiques. Le modèle est caractérisé par une architecture à trois couches (morphologique, sémantique et conceptuelle). Il s'agit de la structure des dictionnaires de la société *Lingway*.

Le chapitre 11 aborde la question de l'indexation automatique des documents. On y présente la distinction entre l'indexation libre et l'indexation contrôlée. Les concepts présentés sont appuyés par des exemples fondés sur la technologie de la société *Lingway* (interrogation en langue naturelle sur le site des *Pages Jaunes* pour l'indexation libre et interrogation de la nomenclature *MedDRA* pour l'indexation contrôlée). Le dernier chapitre de l'ouvrage traite d'un domaine en pleine expansion, celui de l'extraction de données et d'informations à partir de textes. Ce domaine, aussi nommé « forage de textes », vise à développer des méthodes dont l'objectif est d'identifier et de relier certaines informations traitant de sujets précis. Les propos développés dans cette section sont mis en relation avec la problématique de la recherche d'information et appuyés, encore une fois, par des exemples utilisant la technologie de la société *Lingway*.

Dans cet ouvrage, la question de la recherche d'informations est abordée dans une perspective linguistique bien ciblée, radicalement différente de l'approche mathématique que l'on retrouve dans la majorité des ouvrages introductifs traitant du même sujet. Nous nous réjouissons donc de la parution d'un tel ouvrage dans la mesure où il répond à un besoin précis : celui d'introduire les lecteurs au domaine de la recherche d'informations fondée sur le traitement automatique – d'inspiration essentiellement linguistique – de la langue. En outre, il est essentiel de noter que l'ouvrage est très bien rédigé et sa structure, claire et précise, jumelée à de nombreux exercices corrigés, contribue grandement à la compréhension des concepts et des techniques exposés.

Nous regrettons cependant que la bibliographie accompagnant l'ouvrage soit très limitée. En effet, elle n'est composée que de quinze références. Il s'agit d'un choix délibéré et justifié par les auteurs dans l'introduction de l'ouvrage. Comme il s'agit d'un ouvrage dont l'objectif est « avant tout pédagogique » (p. 17), nous sommes d'avis qu'il aurait été souhaitable que la bibliographie soit davantage étoffée afin de guider les lecteurs vers les principales contributions dans le domaine.

En outre, dans la seconde partie de l'ouvrage, certains chapitres reposent en grande partie sur des illustrations employant la technologie de la société *Lingway*. Cette manière de procéder offre aux lecteurs des exemples d'application concrets. Cependant, ces chapitres, bien que pédagogiquement très utiles, pourront, pour certains lecteurs avertis, donner l'impression que les auteurs ont d'abord voulu démontrer l'efficacité de la technologie employée.

Malgré ces deux éléments critiques – mineurs, est-il nécessaire de le souligner – nous recommandons grandement la lecture de cet ouvrage à tous ceux qui souhaiteraient se familiariser avec les concepts essentiels et les techniques fondamentales du traitement de la langue dans son application à la recherche d'informations.

---

**Denis MAUREL, Franz GUENTHNER, Automata and Dictionaries, King's College Publications, 2006, 240 pages, ISBN 1-904-987-32-X.**

**par Aurélie NEVEOL**

*Equipe CISMéF (Rouen) et National Library of Medicine (Bethesda, Etats-Unis)*  
*aurelie.neveol@insa-rouen.fr*

---

*Automata and Dictionaries présente l'utilisation d'automates et de transducteurs comme formalismes de représentation pour des dictionnaires électroniques de langue naturelle. Après un tour d'horizon sur les enjeux de la construction de tels dictionnaires et sur les applications du TAL utilisatrices de ce type de ressources, les auteurs rappellent les définitions formelles et les algorithmes liés à la manipulation des automates et transducteurs. Chaque étape est illustrée par des exemples concrets constitués par des dictionnaires jouets. Cet ouvrage est une bonne introduction technique à la manipulation d'automates dictionnaires et renvoie le lecteur souhaitant approfondir le sujet vers les publications adéquates. On peut cependant regretter que le premier chapitre, plus général, n'ait pas la clarté des suivants.*

Le premier chapitre se fixe pour triple objectif d'exposer les difficultés liées à l'élaboration de dictionnaires pour une langue donnée, les raisons pratiques motivant cette entreprise et les spécificités engendrées par la réalisation de cette tâche au format électronique. Les auteurs partent du constat qu'un flou théorique demeure sur le contenu exact des dictionnaires à élaborer. Ils doivent encoder à la fois les propriétés lexicales des unités de la langue, les liens qui les unissent ainsi que les

règles qui régissent leur association avec d'autres unités – cela, de manière exhaustive. En pratique, les modalités de recueil et de représentation de ces informations ne sont pas unanimement établies : notamment, une étape aussi fondamentale que le découpage des *unités lexicales* pose problème. Cependant, au delà de l'aspect descriptif de la langue, les dictionnaires sont nécessaires dans de nombreuses applications du Traitement Automatique de la Langue telles que l'analyse syntaxique, la recherche d'information, la traduction automatique ou encore la correction orthographique. Ce dernier domaine est celui qui semble avoir le plus profité de l'utilisation de dictionnaires électroniques. Comme pour les autres applications, les dictionnaires disponibles sont néanmoins insuffisants – en termes de couverture et d'informations encodées, pour obtenir les performances souhaitées. Ainsi, il est nécessaire de poursuivre le travail amorcé pour élaborer des dictionnaires électroniques aussi exhaustifs que possible, utilisables par l'ensemble de ces applications. Pour ce faire, il convient d'apprécier l'ampleur de la tâche du fait du nombre de langues à décrire, du nombre de domaines de spécialité à couvrir et de la diversité des applications auxquelles les dictionnaires sont finalement destinés. Ce chantier doit donner lieu à un effort collaboratif rassemblant les chercheurs des différentes communautés scientifiques concernées.

Le deuxième chapitre illustre le fonctionnement et l'intérêt des automates à états finis comme structure de stockage des dictionnaires électroniques. Pour ce faire, un dictionnaire simple composé des jours de la semaine est utilisé. Le troisième chapitre présente des définitions formelles autour des automates et transducteurs. Les notions introduites sont illustrées à l'aide de l'exemple donné au chapitre précédent.

Les chapitres quatre et cinq définissent et explicitent les notions d'automate déterministe et minimal. Les chapitres cinq et six synthétisent les algorithmes de construction des automates et transducteurs introduits par différents chercheurs. Un algorithme particulier est détaillé et implémenté sur un dictionnaire comportant quelques entrées choisies pour illustrer les étapes clés.

Finalement, le chapitre huit aborde des questions liées à la complexité des algorithmes, l'espace de stockage ou le temps d'exécution nécessaire à leur mise en oeuvre. Les chapitres 4 à 7 se terminent par quelques exercices d'application dont les solutions sont données en annexe.

La clarté des chapitres 2 à 8 fait défaut au début de l'ouvrage et l'objectif pédagogique s'en trouve atteint. La longueur des phrases, le style empesé ainsi que l'absence de définitions et de références pour certaines des notions introduites nuisent à la lisibilité de cette partie. Le public novice visé (étudiants) pourra éprouver quelques difficultés à appréhender le sujet traité avec la vision globale souhaitée par les auteurs.

Malgré le cadre volontairement restreint de l'ouvrage annoncé par les auteurs en introduction, certaines ouvertures sur d'autres travaux connexes aux automates et aux dictionnaires électroniques auraient pu avoir leur place. Par exemple, l'affirmation en préface qu'aucune méthode statistique n'a permis de construire de

ressources linguistiques conséquentes<sup>1</sup> porte à controverse à la lumière du travail récemment entrepris à l'INRIA sur l'élaboration du Lexique des Formes Fléchies du Français<sup>2</sup>. Une discussion contrastant les deux approches (essentiellement manuelle vs. statistique) et l'utilisabilité des ressources produites (DELA vs. Lefff) aurait eu un intérêt à la fois pédagogique et pratique. Dans une moindre mesure, une rapide comparaison entre l'utilisation des automates appliqués à des textes en langue naturelle par opposition à d'autres formes de textes telles que les séquences génomiques aurait permis une vue plus complète de la portée des outils décrits.

Dans l'ensemble, malgré quelques écueils, l'ouvrage fournit une présentation simple, claire et illustrée de l'utilisation d'automates et de transducteurs pour la construction de dictionnaires électroniques. Les principes de base de la théorie des automates sont exposés, et nombre de références plus détaillées sont proposées aux lecteurs souhaitant approfondir le sujet. Dans les chapitres 6 et 7, les algorithmes les plus complexes de l'ouvrage sont exposés de manière tout à fait abordable. *Automata and dictionaries* constitue une bonne introduction à la problématique des dictionnaires électroniques sous forme d'automates.

---

**Benoît HABERT, Instruments et ressources électroniques pour le français, Ophrys, 2005, 169 pages, ISBN 2-7080-1119-7.**

par Susanne ALT

ATILF-CNRS  
salt@atilf.fr

---

*Après « Les linguistiques de corpus », B. Habert nous propose, dans la collection « L'essentiel français » des éditions Ophrys, une dissection de la linguistique à l'instrument. Partant de la modification des pratiques d'analyse des données en sciences du langage, l'ouvrage se fixe un double objectif : donner davantage de visibilité aux ressources électroniques (corpus et lexiques) et aux « instruments » (logiciels d'annotation) disponibles pour le français, et combiner cet aperçu avec des réflexions pratiques et méthodologiques sur la mise en œuvre de ces nouveaux moyens de travail.*

Le premier chapitre est consacré entièrement à un exemple concret, le poème « Le dormeur du val » de A. Rimbaud. Ce poème est traité tour à tour par *Cordial*, le *Treetagger*, *Flemm*, *Dérif*, *Syntex*, *Transcriber*, le *Métromètre* et la TEI. Il introduit ainsi de façon pratique toutes les notions importantes du domaine : ressource, instrument, annotation, niveau de description linguistique, langage de balisage,

---

1 p. 8: « [T]here has not been any substantial progress in creating (...) sophisticated linguistic databases (e.g. monolingual (...) lexica (...)) on the basis of statistical approaches (...) »

2 Le Lefff, disponible sous licence libre avec les publications associées sur <http://www.lefff.net>.

format de représentation, métadonnées. Ce premier chapitre constituerait un fondement à la fois empirique et pédagogique extrêmement riche pour la suite, s'il arrivait à passer le cap d'une articulation explicite de ces notions : on aurait souhaité qu'il débouche sur un dessein structurant la suite de la lecture autour des problématiques scientifiques transversales du domaine. Le plan actuel reste malheureusement un peu trop prisonnier d'une présentation des ressources et instruments d'un côté, et de la discussion méthodologique de l'autre.

De fait, les chapitres III à VI répondent au premier objectif annoncé : donner davantage de visibilité aux ressources électroniques et aux instruments. Ils sont plus particulièrement dédiés à la présentation de corpus et outils pour l'étude du « texte », des « paroles » et des « mots ». Sous ces étiquettes (aux linguistes d'en apprécier l'adéquation), le lecteur trouve des informations essentiellement factuelles sur les corpus écrits et oraux, ainsi qu'une section dédiée aux lexiques. Le chapitre VI constitue une ouverture appréciable sur les problèmes particuliers posés par les aspects multimodaux, multilingues et variationnels (diatopiques et diachroniques).

La discussion méthodologique, deuxième objectif annoncé, est répartie sur les chapitres II, VII et VIII, consacrés aux choix d'annotation, questions pratiques et méthodes de la « linguistique instrumentée ». Ces chapitres sont de qualité inégale. Pour une majeure partie, ils proposent des réflexions méthodologiques et théoriques fondamentales pour la représentation, la gestion et l'archivage de ressources linguistiques : propriétés des annotations (embarquées ou débarquées), la question de l'évaluation, les aspects de standardisation, la documentation par des métadonnées. Certaines de ces sections – par exemple les exposés sur l'encodage des caractères ou les métadonnées – sont exemplaires en ce qui concerne l'équilibre entre pertinence thématique, technicité et pédagogie. L'adéquation à la fois pratique et théorique de ces parties contraste d'autant plus étrangement avec certaines sections du chapitre VII : il est par exemple permis de s'interroger sur la contribution effective à un ouvrage sur les ressources linguistiques d'une partie sur l'installation de logiciels (« *si l'instrument existe pour plusieurs systèmes d'exploitation [...], on décharge la version correspondante* »). Dans une moindre mesure, sont également concernées les pages consacrées aux langages de requête propriétaires, aux bases de données relationnelles, aux langages de scripts et aux expressions régulières. Ces pages, trop peu approfondies pour rendre un utilisateur novice opérationnel, ont essentiellement le mérite de traduire le désarroi courant des utilisateurs devant le paysage actuel des ressources électroniques, et d'illustrer l'approche « maison » qui domine pour l'instant leur exploitation effective. Par conséquent, le chapitre VII aurait grandement bénéficié d'être accompagné par un avertissement soulignant le fait qu'il ne peut s'agir là que de solutions provisoires et fondamentalement contraires aux principes de gestion de ressources préconisés par ailleurs dans l'ouvrage (formats semi-structurés, standardisation, documentation par métadonnées) : de facto, tout ce chapitre constitue un excellent argument en faveur du développement d'environnements partagés, ouverts et standardisés, comme par exemple la plate-

forme GATE<sup>3</sup> – exemplaire pour sa modularité et ses interfaces standardisées, ou l’initiative DOBES<sup>4</sup> – exemplaire pour sa gestion de métadonnées linguistiques.

Pour résumer, la dichotomie entre « ressources et instruments » et « méthodologie » a tendance à favoriser la description de l’existant et du provisoire au détriment d’une vision des enjeux scientifiques et technologiques à venir, et brouille quelquefois la clarté du propos. En particulier, certaines thématiques transversales et centrales du domaine se retrouvent éclatées sur plusieurs chapitres et font l’objet de redites : la normalisation (à ce propos, aucune mention du TC 37/SC 4 ?), l’interfaçage des ressources avec des outils, ou la constitution de corpus. Si l’ouvrage représente une bonne introduction aux ressources, outils et pratiques existantes, il n’est pas véritablement convaincant – tout simplement, parce qu’il ne cherche pas vraiment à convaincre. A certains égards, il traduirait presque un constat d’impuissance devant les résultats hétérogènes de l’évolution inéluctablement électronique du français. Pourtant, cette évolution fait émerger de nouveaux enjeux scientifiques et techniques à l’intersection de la linguistique, de l’informatique et des sciences de l’information. C’est cette voie qui mérite d’être creusée, et pourquoi pas en suivant l’invitation à l’action, formulée par B. Habert à la page 10 de l’ouvrage : « *La linguistique se trouve désormais en mesure de recourir à de nouveaux instruments et à des données renouvelées. Il lui faut, dans le même mouvement, en mesurer la nature et l’adéquation à ses objectifs propres, quitte à intervenir pour infléchir l’usage de ces moyens, voire leur conception même.* »

---

**Patrice ENJALBERT, Sémantique et traitement automatique du langage naturel, Hermès-Lavoisier, 2005, 410 pages, ISBN 2-7462-1126-2.**

**par Pascal AMSILI**

*Université Paris 7 & Lattice*  
amsili@linguist.jussieu.fr

---

*Cet ouvrage est un recueil de 10 articles, dont plus de la moitié co-écrits par Patrice Enjalbert. L’objectif est double : d’une part présenter les travaux menés depuis 15 ans à Caen sur la question du sens en TAL, et d’autre part illustrer l’actualité des recherches sémantiques en TAL, et proposer méthodologie et état de l’art sur cet aspect.*

### **Sémantique et TAL, quel rapport ?**

La question peut paraître saugrenue, mais elle mérite qu’on s’y arrête. On peut dire que le TAL, en tant que domaine de recherche, ne se pose en général pas la

---

<sup>3</sup> <http://gate.ac.uk/>

<sup>4</sup> <http://www.mpi.nl/DOBES/>

question dans les termes de la tripartition habituelle entre *syntaxe* (relation des signes linguistiques entre eux), *sémantique* (relation des signes à leur dénotation) et *pragmatique* (relation des signes à leurs utilisateurs). En effet, le TAL est défini essentiellement par sa dimension applicative : il s'agit de mettre au point outils et méthodes permettant de réaliser des traitements de matériau linguistique, lequel est dans tous les cas porteur de sens (et c'est précisément parce qu'il y a du sens que nous utilisons des applications de TAL). Bien entendu, les traitements envisagés travaillent par définition sur la *forme* du matériau (signal pour la parole, chaîne de caractères pour l'écrit). Il n'y a donc pas d'application de TAL (au sens où nous venons de le définir) qui échappe à la présence de ces trois niveaux : un correcteur orthographique, par exemple, dont on peut penser qu'il s'intéresse essentiellement à la surface (au sens linguistique), ne peut faire abstraction de la sémantique (quand, par exemple, la désambiguïsation du sens en contexte permet de désambiguïser syntaxiquement, et ainsi de résoudre un problème d'accord), ni de la pragmatique (puisque la plupart des correcteurs tentent d'intégrer des règles dites d'usage). On peut remarquer d'ailleurs que le TAL ne s'intéresse pas davantage à la syntaxe, ou du moins ne s'intéresse pas à la syntaxe comme fin: déterminer la structure syntaxique d'une chaîne de mots peut être utile pour mener à bien certaines applications, mais, le problème étant notoirement difficile, nombre d'applications se contentent d'approximations sur cette structure syntaxique, ce qui est parfaitement justifié si le traitement visé en est rendu suffisamment efficace.

D'un point de vue différent, on peut considérer la tripartition comme une sorte de guide méthodologique pour élaborer l'architecture d'une application de TAL. On aurait un module syntaxique, un module sémantique, et un module pragmatique. Mais la plupart des applications de TAL ont une architecture tout autre, et certains modules classiques peuvent être vus comme relevant de plusieurs de ces niveaux (l'étiquetage, par exemple, plutôt (morpho-)syntaxique, demande parfois une désambiguïsation des sens en contexte).

A ce point, on pourrait conclure comme le titre provocateur de ce compte-rendu de lecture le suggère, qu'il n'y a pas de pertinence à tenter de rapprocher les deux termes.

Cependant, et c'est ce que montre l'ouvrage dirigé par Patrice Enjalbert, la question devient pertinente si l'on adopte le bon point de vue. D'une part, on peut noter que le TAL puise largement dans les résultats de la linguistique (formelle)<sup>5</sup>. Il est donc naturel de s'interroger sur la façon dont les recherches en sémantique peuvent inspirer les méthodes et les algorithmes du TAL. D'autre part, on peut ouvrir les perspectives, en s'inscrivant non plus dans le TAL au sens étroit, mais dans ce qu'on pourrait appeler la linguistique informatique, ou computationnelle. Si tant est qu'un tel domaine de recherche existe et se distingue de celui de la linguistique formelle, on peut le caractériser de la façon suivante : l'objet d'étude est la langue, et

<sup>5</sup> Bien entendu, il existe des approches du TAL qui ne sont pas linguistiquement inspirées (certaines méthodes probabilistes ou non symboliques). Elles ne sont pas moins légitimes, mais la question traitée ici ne les concerne pas.

les méthodes d'investigation habituelles de la linguistique sont doublées d'un souci de formalisation implémentable. Selon ce point de vue, il devient tout à fait pertinent de s'intéresser à la sémantique. En effet, la sémantique formelle (qu'elle soit lexicale ou non lexicale) élabore des modèles qui peuvent être implémentés, ce qui facilite leur mise à l'épreuve des données. L'ouvrage proposé me semble s'inscrire pleinement dans cette perspective. Il comporte à la fois des discussions sur l'implémentation de modèles linguistiques (2e partie) et des questions sur la façon de « faire entrer de la sémantique » dans des applications standard de TAL (3e partie). C'est la raison pour laquelle il me semble que « sémantique et linguistique informatique » aurait pu faire un meilleur titre. Nous avons donc affaire à un projet pertinent et cohérent.

Ce projet est cependant extrêmement vaste : d'une part, le nombre de modèles sémantiques potentiellement implémentables est important, d'autre part, comme nous le disions plus haut, la sémantique est pertinente à tous les étages des applications TAL. C'est sans doute ce qui explique l'autre parti pris de cet ouvrage : il s'agit d'un ouvrage centré autour des travaux menés à Caen (et à Paris) par une petite équipe de chercheurs, autour de Patrice Enjalbert, Bernard Victorri et Laurent Gosselin. Ce parti pris explique le caractère partiel de l'ouvrage, mais donne aussi une certaine cohérence à l'ensemble.

### Résumé des chapitres

L'ouvrage est organisé trois grandes parties : la **première partie**, intitulée "Repères", est destinée aux débutants en sémantique. Le point de vue adopté n'y est pas linguistique, mais déjà orienté vers les problématiques traitées dans la suite. Le chapitre 1 ("sémantique et TALN, première approche") tente de mener une réflexion, sans a priori théorique, sur ce que peut être la construction du sens (et comment on pourrait la simuler). Le chapitre est basé sur de nombreux exemples de textes variés, à propos desquels est menée une réflexion linguistique sommaire pour illustrer la problématique. Le chapitre 2 ("Les paliers de la sémantique") vise à raffiner la réflexion ébauchée au chapitre 1, en considérant successivement le mot, la phrase et le texte. Comme le précédent, ce chapitre est destiné à des débutants en sémantique, et même en linguistique, et on y trouve des définitions de la morphologie, ou des relations classiques de sémantique lexicale, etc.

La **deuxième partie** s'intitule "Modélisation sémantique". Elle comprend quatre chapitres, qui sont chacun organisés autour du modèle d'un phénomène linguistique, et de sa mise en oeuvre dans un système informatique. Le premier chapitre (ch. 3, "polysémie lexicale") est consacré à un modèle de la polysémie lexicale, proposé par Bernard Victorri, et implémenté. La thèse défendue est que les sens des mots peuvent être représentés dans un espace multidimensionnel, espace que l'on peut construire en utilisant la relation de synonymie. L'objectif principal du modèle est d'ordre linguistique : il est de permettre la visualisation des positions des mots dans

cet espace. Ce chapitre montre aussi comment un tel modèle mathématique peut être utilisé pour une tâche de TAL, la désambiguïsation des sens en contexte. Il s'agit donc avant tout d'une prise de position linguistique, voire cognitive, concernant le lexique et son organisation, et en second lieu de la façon d'utiliser le modèle en TAL. Le deuxième chapitre (ch. 4, "Calcul de la référence") est quant à lui centré sur une application de TAL, en l'occurrence la résolution automatique des anaphores. Il s'agit du système Calcoref, développé par Michel Dupont. La réalisation de ce système a conduit M. Dupont à élaborer une sorte de modèle, largement inspiré de la notion d'accessibilité proposée par Mira Ariel. Dans le troisième chapitre (ch 5, "Temporalité"), c'est clairement le modèle qui est premier : il s'agit du modèle linguistique de la temporalité proposé par Laurent Gosselin. Les ambitions calculatoires de ce modèle, qui vise à prédire les propriétés aspectuo-temporelles des procès décrits dans un texte, permettent d'envisager une implémentation, laquelle a été réalisée par Cédric Person, co-auteur avec L. Gosselin de ce chapitre. On est clairement ici dans le cadre de l'informatique linguistique : le système construit ne vise pas la réalisation robuste et automatique d'une tâche de TAL particulière, mais plutôt la validation "expérimentale" de la théorie de Gosselin. Le dernier chapitre de cette partie (ch 6, "Sémantique de l'espace et du déplacement"), se situe à une position intermédiaire: il ne présente pas de modèle de la spatialité, mais formule à partir d'observations linguistiques simples un "cahier des charges" que devrait respecter un tel modèle. L'auteur, Yann Mathet, dégage les grands paradigmes de relations spatio-temporelles et en propose une mathématisation (qui a fait l'objet d'une implémentation).

La **troisième partie**, "De la compréhension automatique aux applications documentaires", est d'inspiration plus pratique. Il s'agit essentiellement de passer en revue diverses tâches courantes du TAL, et de montrer ce que peut être, dans ces tâches, un composant sémantique.

Conclusion: l'ouvrage dirigé par Patrice Enjalbert présente d'une façon cohérente les efforts menés depuis une quinzaine d'année par une équipe centrée à Caen, et qui a poursuivi avec cohérence et persévérance la prise en considération du sens dans les applications de TAL. L'expérience menée est intéressante et peut inspirer les chercheurs en TAL intéressés par ce point de vue.

---

**Mark STEVENSON, Word Sense Disambiguation: The Case for Combinations of Knowledge Sources, CSLI Publications, 2002, 175 pages, ISBN 1575863898.**

**par Maud EHRMANN**

*Centre de recherche Xerox de Grenoble*  
*Maud.Ehrmann@xrce.xerox.com*

---

*Dans cet ouvrage, Mark Stevenson rend compte de ses travaux sur la désambiguïsation lexicale automatique, amorcés durant sa thèse et enrichis de développements et réflexions ultérieurs. L'auteur défend l'hypothèse selon laquelle l'utilisation combinée de différentes sources et types d'information est un moyen de parvenir à de bons résultats dans un processus de désambiguïsation. Donnant un aperçu général de la discipline, il porte davantage son attention sur les ressources lexicographiques et décrit avec précision le système qu'il a implémenté.*

Inspiré d'un travail de thèse, l'exposé s'organise en trois temps relativement classiques : introduction de la problématique et état de l'art, présentation de l'approche adoptée et du système implémenté, et évaluation.

L'introduction et le chapitre inaugural donnent un aperçu rapide mais efficace de la tâche de désambiguïsation lexicale automatique. Après une courte définition, l'auteur inventorie les différentes applications en traitement automatique du langage (Tal) pour lesquelles la désambiguïsation présente un intérêt, principalement la traduction automatique et la recherche d'information. Il fait ensuite un historique de la "discipline", avant de rendre compte des différentes méthodes pouvant être mises en œuvre (exploitant des ressources de type dictionnaire ou thésaurus, ou exploitant l'information contenue dans les corpus, ou les deux). Tout ceci est exposé de manière efficace, l'auteur pointe les principaux travaux à connaître et éclaircit ses choix terminologiques. Il achève ce chapitre en posant le problème de l'évaluation des systèmes de désambiguïsation lexicale automatique.

Qui dit *désambiguïsation* dit *sens*, et qui dit *lexique* dit *ressource dictionnaire*. Le chapitre 3 questionne ainsi tout naturellement les deux corollaires théorique et méthodologique de la *désambiguïsation lexicale*, à savoir la notion de *sens* et sa représentation au travers de *dictionnaires*. "Meaning and Lexicon" s'intéresse en effet au sens, non en tant que tel mais tel qu'il est traité en lexicographie et, partant, au rôle des ressources de type dictionnaire en Tal et en désambiguïsation lexicale. La question sous-jacente à laquelle tente de répondre ce chapitre est donc "Que représente le sens des dictionnaires et est-il possible d'en exploiter l'information en Tal"? La réponse s'articule en trois mouvements.

Stevenson commence dans un premier temps par décrire les traditionnelles distinctions sémantiques, évoquant, mais sans entrer plus avant dans le débat, les discussions existant à ce sujet. L'auteur ne cherche pas à répondre de façon

systematique aux problèmes posés par la notion de sens mais observe comment cette dernière est traitée dans les dictionnaires, retraçant les évolutions d'une tradition lexicographique séculaire. Les méthodes de conception de dictionnaires ont connu d'importants changements, et la récente exploitation de ces derniers par le Tal pose la question de leur adéquation à cette tâche. Ce problème se pose à deux égards : le type d'information présente dans le dictionnaire d'une part, et son encodage numérique d'autre part. Stevenson suggère qu'un dictionnaire peut être exploité comme ressource lexicale par le Tal s'il possède les qualités suivantes : une large couverture de la langue, des indications linguistiques basiques, et une orientation descriptive plus que prescriptive. Rendre exploitable informatiquement cette information représente un lourd travail, auquel se sont attelés plusieurs éditeurs. Le seul bémol dans cette évolution est la disparition quasi totale des noms propres hors des dictionnaires encyclopédiques, alors qu'ils étaient présents dans les premiers recensements lexicaux et pourraient être utiles dans certaines tâches Tal.

L'auteur se penche dans un second temps sur la question de la conception de ressource dictionnaire (comment collecter et structurer l'information lexicale). Sont tout d'abord présentées trois ressources lexicales, non "prévues" pour le Tal à l'origine, mais largement exploitées comme telles à l'heure actuelle : le *Longman Dictionary of Contemporary English* (LDOCE), le *Roget's Thesaurus* et *WorNet*. On apprend ici l'essentiel concernant ces ressources : leurs caractéristiques ainsi que la notion de sens au cœur de chacune d'elles sont bien détaillées et une étude comparative en donne un bon aperçu général. Le processus lexicographique lui-même est ensuite examiné, et deux études sont présentées à ce sujet, conduisant toutes deux à la conclusion que les lexicographes opèrent des distinctions sémantiques plus fines que les usagers de la langue. Il est observé également que le processus lexicographique peut être automatisé (concordanciers) mais ne peut se passer entièrement du cerveau humain. Enfin, trois critiques de l'exploitation de dictionnaires en tant que ressources pour le Tal sont présentées (Kilgariff, Pustejovsky et Kay), remettant en cause l'efficacité du modèle sémantico-lexical sous-jacent à ce type de ressource au regard d'objectifs tels que la compréhension et la désambiguïsation automatiques.

Quoi qu'il en soit, Stevenson insiste sur le fait que ces ressources ont le mérite d'exister et qu'il importe d'en tenter l'exploitation avant de les dénigrer. Il pointe trois problèmes les concernant et offre des pistes de recherche pour les résoudre. Tout d'abord le fait que les sens soient présentés de manière indépendante : il est difficile de savoir si un sens est proche d'un autre et dans quelle mesure ; la solution à ce défaut peut être le regroupement de sens ou *clustering*. Ces ressources ont aussi pour "inconvenient" d'être statiques ; la solution pour suivre les évolutions de la langue est d'actualiser les ressources (*lexical tuning*) en se basant sur l'analyse de corpus. Enfin, les ressources montrent plus ou moins de spécificités et leur utilisation conjointe, dans un algorithme de désambiguïsation par exemple, doit donc être précédée d'une mise en correspondance des entrées et des informations (*lexical mapping*).

Ce chapitre volumineux met donc en perspective la notion de sens avec les pratiques lexicographiques, insistant plus particulièrement sur les ressources électroniques. Ces dernières sont décrites objectivement et la parole est donnée à leurs détracteurs. Stevenson conclut sur leur probable efficacité quand à la désambiguïsation lexicale automatique.

Faisant suite à l'étude des ressources pouvant être utiles dans une tâche de désambiguïsation, le chapitre 4 propose la définition d'un cadre méthodologique pour leur exploitation. Les ressources lexicales de type thésaurus ou dictionnaires contiennent, pour chaque mot, des informations de natures différentes (syntaxique, sémantique, pragmatique), ces dernières pouvant jouer indépendamment ou conjointement au cours d'un processus de désambiguïsation. Stevenson argumente donc dans le sens d'une exploitation combinée de différentes sources d'information, au travers d'un cadre méthodologique rigoureux dont il décrit ici précisément les tenants et aboutissants (postulats à la base de l'algorithme et types de modules utilisés). Ce cadre est ici "mis à l'essai" avec deux ressources (LDOCE et WN).

Le chapitre 5 concentre son attention sur le rôle de l'étiquetage syntaxique dans un processus de désambiguïsation. Deux expériences sont reportées, montrant toutes deux l'utilité de l'exploitation de l'information syntaxique dans cette tâche : une étude manuelle approfondie de quelques entrées du LDOCE révèle que la plupart des mots peuvent potentiellement, au moins au niveau homographique, être désambiguïsés grâce à la catégorie syntaxique. Cette observation est ensuite validée par une expérience. La prise en compte de ce type d'information passe bien sûr par l'utilisation d'un étiqueteur et les résultats dépendent de la qualité de ce dernier.

Le chapitre suivant rend compte de l'implémentation, suivant le cadre présenté auparavant, d'un système de désambiguïsation exploitant l'information du LDOCE. Le système comporte trois phases : prétraitement, mise en œuvre de modules de désambiguïsation et enfin combinaison de ces informations. Ces étapes sont précisément détaillées et les choix opérés sont justifiés.

Les deux derniers chapitres se concentrent sur l'évaluation, considérant pour l'un la problématique générale de l'évaluation et, pour l'autre, l'évaluation proprement dite du système présenté par l'auteur, selon deux modalités (évaluation générale et étude comparative). L'intérêt de cette dernière, outre le fait qu'elle atteste de la qualité des résultats obtenus par le système, est qu'elle est également mise en œuvre pour chaque source d'information prise séparément, permettant de se rendre compte de la contribution de chacune d'elle au processus général. L'utilisation conjointe des sources d'information reste néanmoins la solution qui présente les meilleurs résultats, allant dans le sens de la proposition principale de l'auteur.

La lecture de ce livre ne peut manquer de satisfaire qui s'y attelle. Celle-ci vaut tout d'abord pour la cohérence et la grande lisibilité de l'exposé. Ensuite, le point de vue général adopté lors de la présentation de la tâche de désambiguïsation (chap.2), des ressources lexicographiques (chap.3) et du problème de l'évaluation (chap. 7 et 8) peut être fort utile au lecteur peu familier du sujet ou cherchant à

conforter ses connaissances dans le domaine. Une critique néanmoins à cet endroit : le propos du chapitre 3 aurait pu connaître un développement moindre, au profit des chapitres 4 et 6 constituant le coeur de l'ouvrage. Enfin, l'axe de recherche proposé, la combinaison de différentes sources d'informations pour la désambiguïsation, selon une démarche d'unification des approches précédentes, est ici servi par une argumentation rigoureuse, la proposition d'un cadre méthodologique et des résultats prometteurs.