Documentation et bibliothèques

Lallich-Boidin, Geneviève et Dominique Maret, *Recherche d'information et traitement de la langue*. 2005. Villeurbanne : Presses de l'Enssib

Lyne Da Sylva

Volume 52, numéro 3, juillet–septembre 2006

URI : https://id.erudit.org/iderudit/1029493ar
DOI : https://doi.org/10.7202/1029493ar

Aller au sommaire du numéro

Éditeur(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (imprimé)
2291-8949 (numérique)

Découvrir la revue

Citer ce compte rendu

Da Sylva, L. (2006). Compte rendu de [Lallich-Boidin, Geneviève et Dominique Maret, *Recherche d'information et traitement de la langue*. 2005. Villeurbanne : Presses de l'Enssib]. *Documentation et bibliothèques*, *52*(3), 218–220. https://doi.org/10.7202/1029493ar

Si l'on fait abstraction des articles encadrés, le dictionnaire est-il si différent des autres éditions du *Grand Robert de la langue française?* Côté format et typographie, il ressemble beaucoup à la deuxième édition « nouveau format » de 2001 en six volumes. Avec ses quatre volumes et son format relativement réduit, cet usuel est donc plus pratique d'utilisation que les autres éditions du *Robert* en six ou neuf volumes. Côté exhaustivité, on y perd un peu: 5 000 entrées de moins que dans le *Grand Robert* (70 000 contre 75 000 et 60 000 dans le *Petit Robert*). Selon toute apparence et logique, ce sont surtout les mots désuets qui ont été sacrifiés. Par ailleurs, il est relativement facile de retracer ces mots dans Internet. Dans sa forme rédactionnelle, ce dictionnaire est un calque mis à jour de ses prédécesseurs. « *Le lecteur va découvrir ici un texte inscrit dans la tradition des ouvrages publiés par* Le Robert, *avec nombre d'enrichissements et de transformations: étymologies développées et précisées, définitions revisitées, descriptions terminologiques mises à jour* » (Préf., p. XV).

Outre le « bon usage », des usages tout à fait récents sont introduits. Le dictionnaire se veut moderne et actuel. À titre d'exemple, mentionnons les entrées ALTERMONDIALISTE, MONDIALISATION, INTERNET, INTERNAUTE, INTRANET, TOILE (la), WEB, WEBCAM, WEBMESTRE, CHAT (clavardage ou bavardage sur Internet). Curieusement, la graphie francisée de « chat » se trouve sous TCHATCHE. L'apport québécois est mentionné et reconnu, notamment pour COURRIEL et TOILE (pour web). Il est à noter que ces mots apparaissaient aussi dans l'édition 2001; les définitions ont cependant été révisées. Par contre, il est étonnant de constater la disparition du mot IMPARTITION. Dans un tout autre domaine, une nouvelle entrée pour une réalité ancienne: ÉVERGÈTE, qui n'était pas dans les *Robert* antérieurs. En ce qui concerne les usages hors de France, les québécismes semblent avoir leur place, du moins POUTINE et SLOCHE; mais on les rattache au français du Canada, sans plus de précision.

Comme dernière particularité, soulignons l'importance des annexes qui font 356 pages à la fin du quatrième volume: Conjugaisons (pp.III-LXXXIV), Petit dictionnaire des suffixes du français (pp. LXXXV-IC), Bibliographies (pp. IC-CCCLVI). La partie bibliographique est plus élaborée que dans les autres *Robert*. On y trouve notamment une bibliographie consacrée aux encadrés culturels.

En somme, ce dictionnaire est un ouvrage novateur et de grande qualité. C'est à la fois un usuel qu'on peut consulter rapidement et un guide pour approfondir un concept.

Jean-Luc FORTIN
Retraité de la Bibliothèque
de l'Assemblée nationale

Lallich-Boidin, Geneviève et Dominique Maret, ***Recherche d'information et traitement de la langue.*** 2005. Villeurbanne: Presses de l'Enssib.

Ce livre arrive à point et comble une lacune importante: une présentation détaillée et ciblée des aspects linguistiques impliqués dans la recherche d'information qui peuvent être exploités par des logiciels. C'est une introduction aux questions pertinentes de traitement automatique de la langue (TAL) dans le but de les utiliser à l'intérieur d'un moteur de recherche.

Structure de l'ouvrage

Tel qu'expliqué dans l'introduction, le livre se présente en deux parties. « *La première partie, plus théorique, traite des fondements de l'analyse linguistique à des fins de traitements automatiques, tandis que la deuxième illustre ces méthodes dans le cadre d'applications à la recherche d'information* » (p. 17). La première partie contient six chapitres, qui traitent des questions suivantes:

- Chapitre 1 — la segmentation d'un texte (segmentation du texte en phrases et des phrases en mots), qui est essentielle et préalable à tout autre traitement;
- Chapitre 2 — la lemmatisation (ou la reconnaissance des flexions du pluriel, du féminin, des terminaisons verbales afin d'identifier le lemme ou la forme de base du mot), qui permet de rassembler toutes les variantes d'un même mot;
- Chapitre 3 — la syntaxe (ou l'étude des relations entre les mots dans la phrase), qui sert à identifier les différents rôles des groupes de mots;
- Chapitre 4 — la sémantique (l'étude du sens véhiculé par les mots et les constructions syntaxiques) et la pragmatique (l'étude de l'utilisation de la langue en contexte), qui sont nécessaires pour véritablement cerner une réponse adéquate à une requête de recherche d'information;
- Chapitre 5 — l'affixation (ou les phénomènes de composition morphologique), qui rend compte de la structure interne des mots complexes, contenant des préfixes et suffixes, ou des mots composés;
- Chapitre 6 — la terminologie (ou l'étude des langues de spécialité), qui tente de capter le vocabulaire spécialisé qui est le plus intéressant pour la recherche d'information, mais qui défie les simples descriptions linguistiques formelles en exigeant des interprétations sémantiques plus ou moins sophistiquées.

La première partie est suivie d'un chapitre sur les « grammaires de réécriture » : ce chapitre *« dont l'objet est plus théorique puisqu'il s'agit de quelques formalismes en vigueur pour le traitement automatique de la langue, pourrait faire l'objet d'une annexe. Sa lecture n'est pas indispensable à la compréhension de l'ensemble »* (p. 17). Il est superflu dans l'ouvrage. Seuls les mordus voudront peut-être s'y aventurer...

La deuxième partie, contenant cinq chapitres, couvre trois thèmes :

- ▷ Chapitres 8 et 9 - la recherche d'information au sens strict, où l'on examine les transformations linguistiques qui peuvent être opérées sur les requêtes en langue naturelle avant de les soumettre au moteur de recherche ;
- ▷ Chapitre 10 — les ressources linguistiques monolingues ou multilingues ;
- ▷ Chapitres 11 et 12 — quelques applications connexes, notamment l'indexation automatique et l'extraction de connaissances.

Partie I — Aspects linguistiques

Les explications linguistiques sont très détaillées, notamment dans le chapitre sur la lemmatisation. Des exemples clairement identifiés illustrent plusieurs des concepts ou procédés présentés. Il y a aussi des exercices utiles pour vérifier la compréhension (les réponses sont données). Les motivations pratiques des propriétés linguistiques sont habituellement présentées en début de chapitre, bien qu'elles soient parfois moins présentes dans le corps du texte. Elles gagneraient à être davantage mises en valeur, et même à faire l'objet d'un récapitulatif à la fin des chapitres, car le lecteur peut facilement se perdre dans les explications théoriques.

Partie II — Applications en recherche d'information

Les chapitres 8 et 9, présentent la description des deux aspects linguistiques de l'appariement entre question et index, c'est-à-dire le traitement des documents, d'une part, et celui de la question, d'autre part. Le chapitre 9 « Recherche d'information en langue naturelle », est important. Il comprend 50 pages, soit presque 20 % de l'ouvrage. Ceci s'explique bien sûr par la thématique du livre.

La présentation de l'étiquetage morphosyntaxique statistique (au chapitre 9, p. 194 et suivantes) est complètement dissociée de celle intégrée dans le cadre de la lemmatisation (au chapitre 2, p. 61 et suivantes), ce qui est très déroutant. Il s'agit en effet de deux approches informatiques différentes pour qui tente de dériver le même type d'information. La section sur les recherches dites « *cross-language* », que l'on aurait voulu traduire par « translinguistiques », couvre un aspect intéressant qui suscite un intérêt croissant à l'heure actuelle.

Dans cette deuxième partie, on commente certaines applications commerciales existantes et l'on présente des exemples concrets de situations de recherche d'information :

- ▷ recherche dans une base d'informations musicales par le nom d'un compositeur ;
- ▷ recherche de marques de commerce dans une base de l'Institut de la Propriété Industrielle (INPI) ;
- ▷ recherche de brevets dans les bases de l'INPI, avec prise en compte du multilinguisme ;
- ▷ l'indexation de rubriques des Pages Jaunes ;
- ▷ l'indexation assistée par ordinateur des effets indésirables de médicaments, à l'aide de la nomenclature MedDRA ;
- ▷ recherche d'information dans une base de données de la presse quotidienne.

Les auteurs utilisent, par exemple au chapitre 10, un système commercial précis (dont il y a pléthore à l'heure actuelle). L'avantage de se concentrer sur un outil en particulier, c'est que la présentation est ancrée dans la réalité, et non dans des exemples d'école. Mais la section sur l'architecture des dictionnaires dans le système Lingway ne semble pas à sa place : elle contient notamment une reprise d'éléments de morphologie, normalement déjà couverts précédemment (aux chapitres 2 et 5). La présentation du niveau conceptuel (pont entre les langues) évoque une approche classique à la traduction automatique par interlingua. On n'y fait cependant pas référence (pas plus d'ailleurs qu'à l'ensemble des travaux en traduction automatique), ce qui est une lacune importante.

Pour l'indexation automatique abordée au chapitre 11, les présentations sont plutôt techniques. Il semble nécessaire de faire les exercices proposés pour bien suivre les exposés. Pour sa part, l'indexation thématique (pp. 242-246) est désavantagée par un exposé imprécis.

Des remarques plus générales

La présentation est claire en règle générale, favorisée par de multiples indices de typographie et de mise en pages. Plusieurs exemples et exercices en font un bon livre d'apprentissage pour une utilisation en salle de classe, par exemple. Cependant, pour une lecture autodidacte certains lecteurs trouveront la matière un peu trop technique. On y retrouve malheureusement un certain nombre de coquilles (y compris des incohérences dans un ou deux exemples, ce qui est plus gênant). Plusieurs chapitres comprennent de

très utiles sections dédiées à une présentation (ou un rappel) des avantages de techniques exposées, ainsi que des problèmes résiduels. Notons enfin que l'index est limité: une seule page, contenant 107 entrées, sans renvois et avec une structuration minimale des entrées, pour ce volume de 274 pages.

Conclusion

Somme toute, il s'agit là d'un très bon ouvrage d'introduction aux questions de traitement automatique de la langue pour la recherche d'information. C'est aussi un bon ouvrage de référence pour les étudiants et les enseignants, mais il peut être considéré trop technique par les praticiens. Par contre, la bibliographie est beaucoup trop limitée pour un ouvrage de ce genre: seulement 15 ouvrages (de plus, sa localisation est inhabituelle, entre le corps du texte et la conclusion).

Des pans importants de travaux en recherche d'information en langue naturelle des 40 dernières années sont occultés, à commencer par ceux de Sparck Jones, ainsi que les multiples travaux plus récents de plusieurs chercheurs de la communauté du TAL (Strzalkowski ou Grefenstette, notamment) et des sciences de l'information (Soergel, Oard, par exemple). De même que chez les auteurs francophones: Gaussier en recherche d'information, Namer en analyse morphologique, etc. ⊙

Lyne Da Sylva
École de bibliothéconomie et des sciences de l'information de l'Université de Montréal

Sources à consulter

Gaussier Eric, Gregory Grefenstette, David Hull et Claude Roux. 2000. Recherche d'information en français et traitement automatique des langues. *T.A.L* 41(2): 473-493

Grefenstette, Gregory. 1998. *Cross-language information retrieval.* Boston: Kluwer Academic Publishers.

Namer, Fiammetta. 2000. Flemm: Un analyseur flexionnel du français à base de règles. *T.A.L* 41(2): 523-548.

Oard, Douglas et Anne Diekema. 1998. Cross-Language Information Retrieval. In *Annual Review of Information Science and Technology* 33, p. 223-256.

Soergel, Dagobert. 1997. Multilingual thesauri in cross-language text and speech retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence.* [http://www.ee.umd.edu/medlab/filter/sss/papers/soergel.ps].

Sparck Jones, Karen, Roger M. Needham et A.H.J. Miller. 1960. *The information retrieval system of the Cambridge Language Research Unit.* Cambridge: Cambridge Language Research Unit, Report ML 109.

Sparck Jones, Karen. 1970. Automatic thesaurus construction and the relation of a thesaurus to indexing terms. *Aslib Proceedings*, 22: 26-28.

———. 2003. Document retrieval: shallow data, deep theories; historical reflections, potential directions. In *Proceedings of the 25th European Conference on Information Retrieval (ECIR-03)*, (réd. F. Sebastiani), Lecture Notes in Computer Science 2633, Berlin: Springer, p. 1-11.

Strzalkowski, Tomek (ed). 1999. *Natural Language Information Retrieval.* Dordrecht: Kluwer Academic Publishers.

Index des annonceurs