

Une source de références : *Infolingua*

Michel LENOBLE

SABOURIN (Conrad F.) : 1994, *Literary Computing: Style Analysis, Author Identification, Text Collation, Literary Criticism: Bibliography* (Montréal-Hudson, Infolingua, « Infolingua », n° 8), 581 pp.

ISBN 2-921173-12-3. Prix : US\$ 80.

SABOURIN (Conrad F.), TARRAB (Elca) : 1994, *Computer Assisted Language Teaching: Teaching Vocabulary, Grammar, Spelling, Writing, Composition, Listening, Speaking, Translation, Foreign Languages; Text Composition Aids, Error Detection and Correction, Readability Analysis: Bibliography* (Montréal-Hudson, Infolingua, « Infolingua », n° 9), 1 066 pp.

Vol. 1 : ISBN 2-921173-13-1, vol. 2 : ISBN 2-921173-14-X. Prix : US\$ 150.

SABOURIN (Conrad F.), LAMARCHE (Rolande M.) : 1994, *Computer Mediated Communication: Computer Conferencing, Electronic Mail, Electronic Publishing, Computer Interviewing, Interactive Text Reading, Group Decision Support Systems, Idea Generation Support Systems, Human-Machine Communication, Multi-Media Communication, Hypertext, Hypermedia, Linguistic Games: Bibliography* (Montréal-Hudson, Infolingua, « Infolingua », n° 10), 862 pp.

Vol. 1 ISBN 2-921173-15-8, vol. 2 ISBN 2-921173-16-6. Prix : US\$ 130.

SABOURIN (Conrad F.), LAMARCHE (Rolande M.) : 1994, *Electronic Document Processing: Document Editing, Formatting, Typesetting, Markup, Storing, Interchanging, Managing: Bibliography* (Montréal-Hudson, Infolingua, « Infolingua », n° 11), 551 pp. ISBN 2-921173-17-4. Prix : US\$ 80.

✉ 3465 Laval; Montréal (Québec); Canada H2X 3C7; E-mail : lenoble@ere.umontreal.ca

SABOURIN (Conrad F.) : 1994, *Quantitative and Statistical Linguistics: Frequencies of Characters, Phonemes, Words, Grammatical Categories, Syntactic Structures; Lexical Richness, Word Collocations, Entropy, Word Length, Sentence Length: Bibliography* (Montréal-Hudson, *Infolingua*, « *Infolingua* », n° 13), 508 pp. ISBN 2-921173-19-0. Prix : US\$ 80.

Dans la collection « *Infolingua* » viennent de paraître, en ce début d'année 1994, dix-sept numéros thématiques de la série de bibliographies extensives et complètement indexées couvrant tous les aspects et les facettes des domaines de recherche se trouvant à l'intersection de l'informatique et des langues naturelles. Les dix-sept numéros, dont huit se composent de deux volumes, ont été compilés pendant 16 ans par Conrad F. Sabourin, consultant spécialisé dans le traitement informatique des langues et des textes.

Les vingt-cinq volumes que compte cette première édition recensent 73 846 entrées bibliographiques, dont 33 % sont des articles de périodiques, 38 % des articles tirés d'actes de colloque, 22 % des rapports de recherche et des monographies et 7 % des thèses et mémoires. Plus de 400 périodiques et plus de 800 actes de colloque ont été systématiquement dépouillés pour produire ces bibliographies.

L'aire géographique des pays d'origine des documents n'est pas uniquement circonscrite à l'Europe et l'Amérique du Nord, mais inclut aussi, dans une moindre proportion, les autres pays du monde. Ainsi, les entrées proviennent à concurrence de 38 % des États-Unis, de 9 % de la France, ainsi que de la Grande-Bretagne et de l'Allemagne. 6 % des données viennent du Japon, 5 % du Canada, 4 % de l'ex-URSS, 3 % de la Scandinavie, 2 % des Pays-Bas et d'Italie, alors que 1 % du matériau bibliographique émane respectivement de Chine, de Taïwan, de l'ex-Tchécoslovaquie et de Belgique. Les 9 % restants ont été collectés dans les autres pays du monde.

La collection « *Infolingua* » couvre les publications rédigées au cours de la période de quarante et quelques années, allant de l'apparition de l'ordinateur à nos jours. Les nombres d'entrées répertoriées par tranche de cinq ans se répartissent comme suit : 1950-1954, 152; 1955-1959, 848; 1960-1964, 2 343; 1965-1969, 4 005; 1970-1974, 5 281; 1975-1979, 8 366; 1980-1984, 14 643; 1985-1989, 25 542 et 1990-1993, 12 138.

Chacun des numéros de la collection « *Infolingua* » est structuré de façon similaire, hormis le cinquième numéro consacré à la génération de texte. Celui-ci comprend également un article introductif à la discipline, dû à Mark T. Maybury. Tous les numéros commencent par une courte préface identique et non spécifique à la thématique traitée. Le lecteur y retrouve des informations générales concernant la collection, les types de sources retenues, la provenance des documents et leur distribution par tranche chronologique. Conrad F. Sabourin y explique ensuite le système d'indexation développé pour permettre une exploitation rapide de ces répertoires bibliographiques. La deuxième partie des ouvrages est constituée par la liste bibliographique en tant que telle. Les entrées y sont triées en ordre alphabétique de nom d'auteur, la date servant de seconde clef de tri. La troisième section des ouvrages, à peu près égale à la seconde en importance, contient l'index des sujets, triés par ordre alphabétique, pouvant présenter une structuration allant jusqu'à quatre niveaux.

Le découpage des 17 numéros est réalisé par regroupement thématique. Dès lors qu'une même entrée bibliographique peut ressortir à plusieurs thématiques distinctes, il est naturel de retrouver environ 15 % des entrées dans plus d'un volume. Cette façon de concevoir les choses a le mérite de ne pas imposer l'achat de plusieurs volumes, là où les domaines et les thématiques des documents se recoupent quelque peu.

La logique présidant à la fabrication de ces bibliographies indexées est la suivante : chaque entrée est numérotée et le numéro est disposé dans la marge de droite, que la page soit paire ou impaire. L'index par sujet renvoie aux références par le biais de ces numéros d'ordre et non par le nom de l'auteur, ce qui eût été moins précis et moins économique. Les mots clefs proviennent d'un thésaurus de plus de 3 800 mots clefs qui a été élaboré aux cours des années de recherche. Les titres indexés au début du projet ont été revus tout récemment afin de normaliser l'indexation en fonction de l'état final du développement du thésaurus. L'indexation porte notamment sur les aspects suivants des publications : la langue traitée, la méthode employée, le type d'application, les concepts linguistiques utilisés, les domaines de recherche, etc.

Les entrées bibliographiques présentent généralement un champ d'information bibliographique par ligne; tant que faire se peut, Sabourin a indiqué les prénoms des auteurs *in extenso*, même dans le cas de prénoms multiples. Le compilateur a toujours essayé d'inclure dans chaque entrée tous les champs nécessaires à son identification exhaustive, quelle qu'en soit sa complexité.

*
* *
*

Les cinq titres de numéros thématiques que je viens de mentionner en début de compte rendu ont été sélectionnés certes en fonction de ma spécialisation, mais aussi et surtout en fonction des intérêts du lectorat de *RISSH*. Pour les domaines qui me sont les plus familiers, comme la littérature computationnelle, la critique littéraire assistée par ordinateur et la littérature générée par ordinateur, force m'est de concéder que les volumes bibliographiques de Conrad Sabourin sont beaucoup plus complets que toutes les compilations bibliographiques qu'il m'a été donné de consulter jusqu'à présent. Dans bien des domaines traités, pour ne pas dire dans tous, ces ouvrages de référence sont de loin les meilleurs sur le marché de l'information scientifique. Tant le nombre, la complétude et la variété des entrées que l'indexation fine impressionnent. Ces publications surpassent de loin les *The Humanities Computing Yearbooks*, puisque ces derniers ne sont pas indexés, mais seulement divisés en grandes catégories et qu'ils ne sont ni aussi systématiques ni aussi exhaustifs.

Les ouvrages compilés par Conrad Sabourin se distinguent aussi par l'ampleur et la diversité des origines géographiques des sources, par l'étendue chronologique (ce qui permet une étude de l'évolution des disciplines), par le nombre de langues et de littératures traitées (plus de 150) et par la multitude des champs de recherche abordés. La réalisation de ces publications, travail herculéen, vient combler le hiatus séparant deux cultures : celle de l'informatique et celle de la langue prise au sens le plus large du terme; hiatus qui rendait tout effort de documentation ardu et complexe. La structure

même des ouvrages permet deux types de lecture : l'une, cursive et l'autre ponctuelle, par le biais des index.

Avec plus de 50 % des titres recensés qui datent d'après 1985, les 17 numéros de la collection « Infolingua » s'adressent directement aux laboratoires de pointe, aux bibliothèques de recherche, aux centres de documentation, aux doctorants et aux chercheurs individuels. Même si le prix des volumes peut paraître à certains un peu élevé, il n'est rien en comparaison de ce que coûterait une recherche visant à établir ne fût-ce qu'une partie d'un seul de ces relevés bibliographiques thématiques.

Les volumes examinés pour ce compte rendu présentent les défauts de leurs qualités, car l'abondance des domaines couverts n'est que trop peu suggérée par les trop courts sous-titres des divers numéros et car le raffinement des index requiert quelque habitude avant que le lecteur n'en perçoive toute la richesse.

Outils fabuleux et uniques en leur genre, les 25 volumes de la collection « Infolingua » sont désormais et pour plusieurs années à venir *la source de référence* pour ceux qui œuvrent en nos domaines. Ils s'imposent comme le point de départ incontournable de tout projet de recherche computationnelle portant sur un aspect des langues naturelles. Je ne puis que souhaiter voir sortir régulièrement les compléments de mise à jour afin de pouvoir suivre aisément l'évolution de ma discipline de spécialisation sans être forcé de dépouiller une multitude de documents variés et souvent difficile à repérer.

*
* * *

Tous les volumes ont été publiés en 1994 et sont distribués par Infolingua Inc., Boîte Postale 187, Snowdon, Montréal (Québec), H3X 3T4 – Canada. E-mail : 73651.2144@compuserve.com

Le compilateur principal en est toujours Conrad F. Sabourin, assisté par des collaborateurs spéciaux pour quelques numéros. Chaque mention des numéros thématiques est accompagnée du nombre d'entrées s'y trouvant. La liste s'établit comme suit :

Computational Morphology: Morphological Analysis and Generation, Lemmatization: Bibliography (Montréal-Hudson, Infolingua, « Infolingua », n° 1), 492 pp., 2 350 références.

Computational Parsing: Syntactic Analysis, Semantic Analysis, Semantic Interpretation, Parsing Algorithms, Parsing Strategies: Bibliography (Montréal-Hudson, Infolingua, « Infolingua », n° 2), 2 volumes, 1 029 pp., 5 180 références.

Computational Lexicology and Lexicography: Dictionaries, Thesauri, Term Banks; Analysis, Transfer and Generation Dictionaries; Machine Readable Dictionaries; Lexical Semantics; Lexicon Grammars: Bibliography (Montréal-Hudson, Infolingua, « Infolingua », n° 3), 2 volumes, 1 031 pp., 5 910 références.

Computational Text Understanding: Natural Language Programming, Argument Analysis: Bibliography (Montréal-Hudson, Infolingua, « Infolingua », n° 4), 657 pp., 3 830 références.

Computational Text Generation: Generation from Data or Linguistic Structure, Text Planning, Sentence Generation, Explanation Generation: Bibliography (Montréal-Hudson, Infolingua, « Infolingua », n° 5), 649 pp., 2 870 références.

- Natural Language Interfaces: Interfaces to Databases, to Expert Systems, to Robots, to Operating Systems, and to Question-Answering Systems: Bibliography* (Montréal-Hudson, Infolingua, «Infolingua», n° 6), 2 volumes, 847 pp., 4 100 références.
- Machine Translation: Aids to Translation, Speech Translation: Bibliography* (Montréal-Hudson, Infolingua, «Infolingua», n° 7), 2 volumes, 1 168 pp., 8 070 références.
- Literary Computing: Style Analysis, Author Identification, Text Collation, Literary Criticism: Bibliography* (Montréal-Hudson, Infolingua, «Infolingua», n° 8), 581 pp., 4 060 références.
- Computer Assisted Language Teaching: Teaching Vocabulary, Grammar, Spelling, Writing, Composition, Listening, Speaking, Translation, Foreign Languages; Text Composition Aids, Error Detection and Correction, Readability Analysis: Bibliography* (Montréal-Hudson, Infolingua, «Infolingua», n° 9), 1 066 pp., 8 010 références.
- Computer Mediated Communication: Computer Conferencing, Electronic Mail, Electronic Publishing, Computer Interviewing, Interactive Text Reading, Group Decision Support Systems, Idea Generation Support Systems, Human-Machine Communication, Multi-Media Communication, Hypertext, Hypermedia, Linguistic Games: Bibliography* (Montréal-Hudson, Infolingua, «Infolingua», n° 10), 2 volumes, 862 pp., 5 680 références.
- Electronic Document Processing: Document Editing, Formatting, Typesetting, Mark-up, Storing, Interchanging, Managing: Bibliography* (Montréal-Hudson, Infolingua, «Infolingua», n° 11), 551 pp., 4 260 références.
- Computational Character Processing: Character Coding, Input, Output, Synthesis, Ordering, Conversion; Text Compression, Encryption, Display; Hashing; Literate Programming: Bibliography* (Montréal-Hudson, Infolingua, «Infolingua», n° 12), 580 pp., 4 120 références.
- Quantitative and Statistical Linguistics: Frequencies of Characters, Phonemes, Words, Grammatical Categories, Syntactic Structures; Lexical Richness, Word Collocations, Entropy, Word Length, Sentence Length: Bibliography* (Montréal-Hudson, Infolingua, «Infolingua», n° 13), 508 pp., 3 100 références.
- Mathematical and Formal Linguistics: Grammar Formalisms, Grammar Testing, Logics, Quantifiers: Bibliography* (Montréal-Hudson, Infolingua, «Infolingua», n° 14), 612 pp., 3 840 références.
- Computational Speech Processing: Speech Analysis, Recognition, Understanding, Compression, Transmission, Coding, Synthesis; Text to Speech Systems, Speech to Tactile Displays, Speaker Identification, Prosody Processing: Bibliography* (Montréal-Hudson, Infolingua, «Infolingua», n° 15), 2 volumes, 1 187 pp., 8 290 références.
- Computational Linguistics in Information Science: Information Retrieval (Full-Text or Conceptual), Automatic Indexing, Text Abstraction, Content Analysis, Information Extraction, Query Languages: Bibliography* (Montréal-Hudson, Infolingua, «Infolingua», n° 16), 2 volumes, 1 047 pp., 6 390 références.
- Optical Character Recognition and Document Segmentation: Character Preprocessing, Thinning, Isolation, Segmentation, Feature Extraction; Cursive and Multi-Font Recognition, Writer/ Scriptor Identification: Bibliography* (Montréal-Hudson, Infolingua, «Infolingua», n° 17), 512 pp., 3 700 références.