However, such commentary can be very well thought out and instructive, indeed so helpful that in the end it is a much better guide to what to read than the reader of the bibliography would be able to figure out on his or her own. In addition, the author complains that the commentary disrupts the reading of the list. This ignores the fact that a hypermedia database could easily display the references in the list the author wishes to see, and require a click by the reader to see the annotation. If only information science professionals existed, they could come to the rescue in such a situation! One last example: the author does not like boolean searching. In 30 years of searching, he has apparently never gotten satisfying results from the use of AND and OR (p. 80). Presumably he's talking about noise, silence, and false drops in full text searching. But further in the text (p. 132), boolean searching becomes a time saver, and it works "because of the logical continuity which governed the indexing of the ensemble of the materials gathered". Isn't that the value that information science people add to information?

It is true that computer science is reinventing information science as it discovers the problems we've been studying for decades. Computer science doesn't know we exist either. Thus what we call classification takes on the name ontologies as computer scientists discover the need for them, cataloguing data become descriptive metadata, and so on. Organising information is very different from organising data. Information scientists, who take organising information to be the focus of their activities, use computers as their main work tools, but this does not mean that they are competing with computer scientists. On the contrary, cooperation is needed more than ever.

There is an evident bias in this work toward methods used in the English-speaking world, especially the USA. There are snide references to the Bibliothèque nationale de France (p. 19, 49), while the Library of Congress catalogue is "the best and richest source of information" (p. 40), "admirable down to the finest detail" (p. 84). In the eyes of the author, Americans do everything so much better. In their handbooks, they provide relationships between bibliographic items by commenting and including "further readings" etc. At least in Québec, the French-speaking academic community includes such literature reviews, "état de la question" and so on in theses, research reports, and many other texts. Perhaps this is due to the influence of North American English speakers rubbing off on them!

There is a plea for rigour in bibliography, but Professor Varet does not discuss back-of-the-book indexing, nor conceptual indexing in the context of the Web. This is surprising since his bias toward the English-speaking world and his concern for rigour in information tools should favour such a discussion. Indexing has in common with bibliography that it is an intellectual activity that adds value to texts and makes them more useful, and like bibliography, it is not just computer output. Furthermore, the English-speaking world is good at it and the French-speaking world is not. It would be interesting to hear the author's reflexions on this related subject.

To his credit, the author adds humourous remarks here and there, which help lighten up the tone of this text that is mostly serious discourse that would be difficult to decode for those outside the field. It is a philosophical reflexion and quite clearly not a work of scholarship about information science, nor does it claim to be. Ultimately, this book can be considered outside the scope of literature for the information science community, except for those few who theorize about the nature and function of information science. For such readers, it is a provocative piece on a number of important information science issues.

James Turner

Dr. James M. Turner, École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128 Succursale Centre-ville, Montréal, H3C 3J7, Canada, e-mail: james.turner@umontreal.ca.

JACQUEMIN, Christian. **Spotting and discovering terms through natural language processing.** Cambridge, MA: MIT Press, 2001. 378 p. ISBN 0-262-10085-1

The book deals with specific experiments in automatic methods of identifying (spotting or discovering) terms in texts. The book's objectives are to show that:

(i) terms (for example, controlled terms from a thesaurus) appear in many variant forms in texts and any method of term spotting which ignores this fact is limited in scope;

(ii) variations can be captured by a simple, linguistically motivated and computationally efficient language and processor (the FASTR system, designed by the author);
(iii) this approach can be applied to many tasks, including thesaurus construction, automated indexing, etc. which all involve "spotting terms".

The central theme of the book is thus to demonstrate that the task of spotting terms in texts cannot rely solely on the identification of a fixed set of seed terms. Terms are linguistic expressions which undergo transformations when they are embedded in continuous text. For example, the controlled term *blood cell* was reported to occur in the following forms in the experimental corpora: *cells in blood* (p. 167), *cells from peripheral blood* (p. 231) and *blood mononuclear cell* (p. 232). The author thus admittedly departs from orthodox ISO definitions of terms. Results from an experiment seeking syntactic and morphosyntactic variants (p. 292) show that 38.5% of terms spotted by FASTR are indeed variants of controlled terms, supporting Jacquemin's claim.

Jacquemin recognizes three types of variants: syntactic variants, where the order of words may change and other words may be added (*cells from peripheral blood* above); morphosyntactic variants, in which a word undergoes a change in its morphological form (*cell component* and *cellular component*, p. 281); and semantic variants, where a term and its variants differ in that, for example, the variant uses a synonym of a word in the term, (*cell death* and *cell destruction*, p. 301).

A second theme, necessary to support the first, is to present the "machinery" devised to recognize variants. FASTR uses metarules, transformations that relate one set of rules (basic controlled term structure rules) to another (possible variants). For example, a "permutation" rule relates terms built of two nouns (*cell structure*) to variants introducing a preposition (*structure of (the) cell*). A grammar of term structure is developed (mostly noun phrase rules), expressed in a unification-based, shallow parsing formalism (where rules only cover short word spans and only some features of the words are subject to unification). Preliminary, overpermissive rules are applied to a training corpus and tuned manually into filtering metarules, by adding constraints to remove most spurious variants. Morphosyntactic variants are captured by similar metarules which make use of morphological information (a shared root between words, for example, *measure* and *measurement*). Semantic variants are described similarly, by calling on existing thesauri: WordNet 1.6 and the thesaurus in Word97.

The experiments are described and results are examined through the lens of qualitative and quantitative factors. Qualitative evaluation involves the identification of correct terms (i.e. those in the controlled terms list) followed by measures of precision, recall and precision of fallout. The quantitative evaluation examines the proportion of spotted terms to the number of words in the corpus, as well as a comparison of the ratio of terms spotted as such or as variants.

A third theme found in this book, scattered among various subsections, is description of the potential applications of this approach. They include automatic indexing and automatic terminology extraction, but also thesaurus enrichment (pp. 221-272), cross-language information retrieval (p. 306), document filtering in Web searches (p. 307), and term clustering for terminology databases (p. 309), among others.

The book is divided into 9 chapters, which can be grouped as follows. The first three chapters are introductory: Chapter 1 presents motivation for the work. Chapter 2 is an overview of previous experiments in term acquisition (i.e. thesaurus or terminological database construction) and automatic indexing; in itself, it is a useful summary of other related work. Chapter 3 is devoted to an informal linguistic description of what terms are (definitions, etc.) and what is required to describe them in an automatic system (questions of morphology, syntactic structure, parsing techniques and formalisms). The FASTR formalism is presented along with the general algorithms used for parsing. Chapters 4 to 8 delve into the practical details: the grammar of metarules, the experiments and the results of the FASTR system. In chapter 4, the author presents the metarule formalism for syntactic variations, and lists the preliminary metarules. Chapter 5 describes the constraints added in order to produce the filtering metarules, and Chapter 6 explains how the metarules can be used to discover additional candidate terms to enrich an existing term list (in short, by extracting the difference between the original term rule and the output metarule). Chapters 7 and 8 deal with morphosyntactic and semantic variants, respectively. For the syntactic and morphosyntactic variants, evaluation measures are presented. Finally, Chapter 9 concludes by summarizing the results of the work. Appendices contain the metarules file (Appendix A), the form of extracted candidate terms (Appendix B), a description of the corpora and term lists used (Ap-

pendix C) and the grammar files (actually, dictionary entries and basic term rules, Appendix D). A short glossary is provided for special terminology introduced by the author. The book also includes an author index and a subject index.

This is essentially a book for a natural language processing (NLP) audience, but it addresses problems of terminology or thesaurus building, thus assuming a good knowledge of the latter. For matters of NLP, it aims to be self-contained in the sense that it presents short introductions to a number of theoretical domains: unification-based language description and parsing; finite-state techniques; morphological description. These cannot be considered thorough presentations, although extensive references are given to basic works in each field.

The book proves an interesting link between unification-based approaches to language description, corpus linguistics and automated term spotting. Computer scientists may consider it too informal while linguists may consider it naïve in its approach to linguistic description. Nonetheless, it is a successful experiment in shallow parsing conjoined with linguistic expertise using a seemingly computationally efficient approach. The applications are clearly many-fold and timely. The approach is similar to current research in shallow parsing of noun phrases but is original in presenting explicit devices for capturing term variants, an important issue.

The book provides, for all aspects of the work, detailed descriptions as well as concise summaries; the honest, objective attitude of the author allows a proper appraisal of its merits. The evaluation measures presented with each experiment give a clear idea of the basis of the overall assessment of the performance. In addition, multiple examples of terms (and incorrectly spotted non-terms) allow the reader to follow the steps involved in designing and refining the rules.

This is not an example of natural language learning, nor of statistical language processing. Basic term rules are devised automatically, but with the use of a knowledge-rich dictionary. Metarules appear to have been determined by linguistic experts (although no details are given); they are further refined by manual observation of the results. The author considers this an advantage:

> As in the case of syntactic variations, the definitions of these classes [of morphosyntactic variations] does not result from introspection, but rather from experiments on large and diverse corpora. The strength of the linguistic data presented in this study is their experimental grounding ... (p. 278-279)

One may consider the wide scope of the book to be its major weakness. The work relies on results from a great number of different sources (corpus linguistics, theoretical linguistics, parsers, unification, statistical processing, shallow parsing, finite-state techniques, terminology, indexing, thesaurus construction, information retrieval), all of which cannot receive a thorough treatment within this book. (Accordingly, the bibliography is large and wide-ranging.) The overlapping vocabulary may be confusing. Also, the reader is sometimes left with the impression of not grasping all the details of the notation. A notable flaw is the use of a complex calculus of rule and metarule description which the author does not define; rather, he refers to previous work.

Lastly, weaknesses in the recognition of morphosyntactic variants (precision = 45,8% and recall = 58,4%, p. 289) are not sufficiently addressed. This results from a naïve approach to morphological variation, which assumes that any morphological variation of a word will represent the same term. The author recognizes specific variations (namely, prefixes) in which this is usually untrue. But this approach is too liberal; there is a sense that many of these "variants" would be considered at a different level by a terminologist or an indexer: not quite occurrences of the same term, but merely suggestive of the controlled term's relevance as a related concept.

One must finally note that the book seems to have been completed in haste, suffering from a number of typographical errors and, at times, non-idiomatic English usage. But overall, this is an important, interesting work that will surely lead to further investigations.

Lyne Da Sylva

Dr. Lyne Da Sylva, École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal QC, H3C 3J7, Canada.
E-mail: Lyne.Da.Sylva@Umontreal.CA