Patterns of Text: In Honour of Michael Hoey

Mike Scott and Geoff Thompson (editors)
(University of Liverpool)

Amsterdam: John Benjamins, 2001,
vii+323 pp; hardbound, ISBN
90-272-2572-9 and 1-55619-792-6,
$100.00, €110.00

Reviewed by
Graeme Hirst
University of Toronto

Patterns of Text is a collection of papers on the structure of text and on lexical repetition within and across texts. The computational importance of the work is mostly indirect; it is a volume in text linguistics and corpus linguistics rather than computational linguistics per se. Although the corpus research is often computer-assisted, the analysis of the data nonetheless relies mostly on human intuition. The papers draw in particular upon the work of Michael Hoey (e.g., 1983, 1991) and of those upon whom he in turn draws, most notably Eugene Winter (e.g., 1982) and M. A. K. Halliday and Ruqaiya Hasan (e.g., Halliday 1994; Halliday and Hasan 1976). (Indeed, *Patterns of Text* is a Festschrift for Hoey; more on this below.)

Most of the authors in the collection are, or have been, associated with the University of Liverpool or the University of Birmingham, which are centers for this research. (Hoey is Baines Professor of English Language at the University of Liverpool and previously worked at the University of Birmingham.) Birmingham, in particular, is the home of the COBUILD project on corpus-based lexicography and of the enormous *Bank of English* corpus, and one of the contributors to the volume is John Sinclair, editor-in-chief of the *Collins COBUILD English Language Dictionary* (1987). Another distinguished contributor, also from Birmingham, is Malcolm Coulthard, who is a major figure in the fields of discourse analysis and forensic linguistics.

Although only a few of the papers in the book have any explicit computational content, the concerns of many of the papers nonetheless mirror those of recent research in computational linguistics—determining the logical structure of a text, dividing a text into segments, detecting evaluative opinions that are explicit or implicit in a text, learning lexical relations from corpora, detecting plagiarism—and so this volume offers a different and useful perspective on these problems. In view of the similarities of interest, the school of study that the volume represents has received surprisingly little attention in mainstream computational linguistics and vice versa. For example, Hoey's (1991) work on lexical repetition and its use in text abridgement is similar in many ways to that in computational linguistics on lexical chains (Morris and Hirst 1991) and their use in text summarization (Barzilay and Elhadad 1999), but one would not discover this from a citation analysis on either side. Similarly, Hoey's (2001) work on text structure is an important complement to rhetorical structure theory (Mann and Thompson 1988), which has been extremely influential in computational linguistics (e.g., Marcu 2000), but again each side hardly acknowledges the existence of the other. And whereas Levin's (1993) book on verb alternations is much cited in computational linguistics, the COBUILD group's complementary work on pattern grammars (Hunston and Francis 2000) has received little attention in the field (but see Johnson's [2001] enthusiastic review of Hunston and Francis [2000] in this journal last year).

The paper in this volume that most explicitly connects with research in computational linguistics and shows the computational use of Hoey's work is that of Tony Berber Sardinha, "Lexical Segments in Text." The paper is a summary of his 1997 dissertation, which was supervised by Hoey. Berber Sardinha presents a method of text segmentation, called *the link set median procedure*, that is based on the sentence links that are implicit in lexical repetition. It is hard to do justice to the subtlety of Berber Sardinha's procedure in a short summary, but in essence, the links of each sentence can be thought of as covering an area of text, and their median as a kind of center of gravity. The procedure looks for discontinuities in the distribution of the medians and hypothesizes that they are segment boundaries. Berber Sardinha compares his method with Morris and Hirst's (1991) lexical chains and Hearst's (1997) TextTiling on a corpus of 300 texts. (Although Morris and Hirst's work, like Hoey's, is founded on that of Halliday and Hasan [1976], Morris and Hirst used a much broader, thesaurus-based definition of a link, and had no notion of link medians. Hearst's procedure, like link sets, considers only lexical repetition but looks for relatively low values of the cosine similarity between blocks of text to determine boundaries.) Berber Sardinha found that the link set median procedure performed better than lexical chains, but not as well as TextTiling.

Hearst also appears as a computational foil in Antoinette Renouf's paper "Lexical Signals of Word Relations." Renouf's goal is to develop automated procedures for extracting sense relations from text by means of text patterns that serve as signals or cues for the relations. For example, *such as* signals the hypernymy relation in *predators such as the badger*. Renouf criticizes a set of such patterns presented by Hearst (1992), claiming that they are insufficient for dealing with the complexities of their usage as seen in text corpora and hence not suitable for blind, automatic use. Renouf offers a manual analysis of corpus examples of several such signals. The editors of the volume underscore the point in their introduction to the paper: "It is not possible to use [corpus-linguistic] techniques without recourse to one's intuitions" (page 36). Indeed, computational linguistics research often exhibits a tension between full automation for production use of an application system, where some degree of error is deemed to be acceptable, and human-in-the-loop (lexicographic-style) work, especially in the development of resources for use in other applications, where error is not acceptable. Nonetheless, the work of Hearst that Renouf criticizes is now 10 years old, and much has been done since then on the automatic or semiautomatic acquisition of hyponymy relations and ontologies from text (e.g., Hearst 1998; Caraballo 1999; Morin and Jacquemin 1999; Maedche and Staab 2000) and, more generally, on the determination of lexical patterns for extracting information from text (e.g., Byrd and Ravin 1999; Thelen and Riloff 2002).

Malcolm Coulthard's paper on the detection of plagiarism begins with an interesting discussion on the distinction between allusion and plagiarism. Coulthard then presents a method for detecting likely plagiarism, in the face of superficial modifications by the plagiarist, by looking for those sentences in one text that contain at least several words from some sentence in the other text. Put this way, the method sounds obvious, but Coulthard obscures it by couching it in terms of Hoey's vocabulary of *links* and *bonds* and Hoey's methods of text abridgement that look for textually similar sentences, while never actually specifying it in sufficiently precise algorithmic terms. Overall, Coulthard's paper is interesting but also anecdotal and frustratingly informal, and rather carelessly written. (Even the title of Coulthard's paper, "Patterns of Lexis on the Surface of Texts," is a cute but unhelpful play on the titles of two books by Hoey, whereas "Detecting Plagiarism" would have told the potential reader what the paper is about.)

While space does not permit a discussion of all the other papers in the book, two more should be mentioned at least briefly.[1] Mike Scott's paper "Mapping Key Words to *Problem* and *Solution*" describes a computational corpus study that looked for words statistically associated with the words *problem* and *solution* with a view to using them in helping to identify problem-solution structures in texts; the results reported, however, are essentially negative. And Susan Hunston's paper "Colligation, Lexis, Pattern, and Text" is an interesting overview of the semantic subtleties and nuances (in her terms, *semantic prosody*) that are associated with a speaker's or writer's choice not just of individual words but also of phrases and patterns. For example, the pattern *see the amount of* signals an unexpectedly large amount (*when you see the amount of money the CEOs of these organizations are making . . .*); the pattern *what follows is*, when sentence initial, frequently signals an evaluation (*What follows is a poignant memoir . . .*).

As indicated by its subtitle and a few remarks at the end of the editors' introduction, *Patterns of Text* is a Festschrift for Michael Hoey. Usually, a Festschrift records the special influence of the subject's career and research program upon his or her field of study. It will therefore include at least a short biography of the subject (and usually a photograph); an overview of his or her research, explaining its importance and its influence upon the work of others; and a bibliography of the subject's publications. None of that is present here. We don't even learn where Hoey works, we get no list of his publications except for those cited by the individual papers, and, notwithstanding the editors' introduction, we learn very little about Hoey's work or why it is distinguished from that of his peers.[2] Rather, the introduction merely mentions his name a few times and cites a couple of his papers (not even his major books), as if he were just a typical one of many researchers on the topic. And although most of the papers cite Hoey's work (Fries and Sinclair don't), the citations sometimes seem peripheral and motivated primarily by the paper's inclusion in this collection.

1 An additional paper that very explicitly addresses computational issues, but not in a helpful way, is that of John Sinclair, entitled "The Deification of Information." Superficially, it might be thought of as making the case for a massive increase in research in computational linguistics and natural language interfaces and hence should be much appreciated by readers of this journal. But it is actually just an embarrassing fulmination against the World Wide Web and the poor quality of user interfaces in general, without ever using the terms *World Wide Web* or *user interface*. Sinclair claims that no user interface for the provision of information (or, to judge from some of his comments, no user interface at all) can be effective unless it permits true "two-way" (i.e., mixed-initiative) conversation in the way that human language does—as if traditional libraries were "conversational" or mixed-initiative. In effect, he says: "I don't get it; therefore it is ungettable; and those people who think they get it are deluded." He backs his argument up with vast, unsubstantiated generalizations and outright absurdities: "The dominant models of communication are not well suited to humans, and deter most of them from full participation in the benefits of the information cornucopia" (page 295); "The argument that people will eventually adapt is persuasive, because they obviously have the capacity to do so, and if no alternative is provided, no doubt many will in time, though they will have to put up with a degraded form of communication compared to what can be achieved using natural language. It will be a hazardous experiment; if it succeeds, the nightmare scenario of human beings being dominated and even ruled by machines will become that much nearer, since there is no doubt that surrendering one's discourse agenda is an act of gross subservience" (page 308). Well, of course, every researcher in the design of user interfaces and user interaction is well aware that enormous problems remain unsolved and that even current knowledge is frequently ignored (Cooper 1999; Johnson 2000). But Sinclair seems to be unaware of most work in computational linguistics on conversation and dialogue, especially mixed-initiative dialogues, and of research in the design of (nonlinguistic) user interfaces and human-computer interaction. The editors concede in their foreword that Sinclair's paper, "though first committed to paper only a few years ago, now looks dated" (page 288), but this is the least of its problems. Its publication is a misjudgment by both the author and the editors.

2 Interestingly, Hoey has edited or coedited Festschriften for two of the contributors to this volume: Sinclair, Hoey, and Fox (1993) for Malcolm Coulthard and Hoey (1993) for John Sinclair. The former contains all the elements mentioned; I was unable to obtain a copy of the latter.

In general, the quality of the writing in the volume is not high, and Renouf's paper and that of Edge and Wharton (on teaching student teachers to write by teaching them to understand text structure) are notably mediocre in this respect. Coulthard's paper presupposes the reader's familiarity with the content of T. S. Eliot's poem *The Waste Land*, which is rather unrealistic for a scientific paper that presumably seeks a wide international audience. The copyediting is mostly competent (with occasional lapses), though there has been little attempt to harmonize the style of the papers in matters such as the presentation and numbering of examples. Sometimes style varies within a single paper; for example, in Renouf's paper, within just two pages (pages 42–43), italics, quotation marks, and upper case are all used as a metalinguistic indicator; and the format of her Table 10 varies without reason from that of her other logically equivalent tables.

The study of patterns in text and the approach that *Patterns of Text* exemplifies are becoming increasingly important in computational linguistics, natural language processing, and their applications, but despite some bright spots, this book is overall a disappointing presentation of the ideas. Instead, readers might wish to turn directly to the work of Hoey (1983, 1991, 2001) and to related work such as that of Hunston and Francis (2000).

Acknowledgments

I am grateful to Hector Levesque, Gerald Penn, Suzanne Stevenson, and Nadia Talent for helpful comments on earlier drafts of this review. Preparation of the review was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Barzilay, Regina and Michael Elhadad. 1999. Using lexical chains for text summarization. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, pages 111–121.

Byrd, Roy and Yael Ravin. 1999. Identifying and extracting relations in text. In *Proceedings of the Fourth International Conference on Applications of Natural Language to Information Systems (NLDB-99)*, Klagenfurt, Austria.

Caraballo, Sharon A. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 120–126, College Park, MD.

Cooper, Alan. 1999. *The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity*. Sams Publishing, Indianapolis.

Halliday, M. A. K. 1994. *An Introduction to Functional Grammar*. Arnold, London, second edition.

Halliday, M. A. K. and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 539–545. Nantes, France.

Hearst, Marti A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Hearst, Marti A. 1998. Automated discovery of WordNet relations. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press, pages 131–151.

Hoey, Michael. 1983. *On the Surface of Discourse*. Allen and Unwin, London.

Hoey, Michael. 1991. *Patterns of Lexis in Text*. Oxford University Press.

Hoey, Michael. 1993. *Data, Description, Discourse: Papers on the English Language in Honour of John McH Sinclair*. HarperCollins, London.

Hoey, Michael. 2001. *Textual Interaction: An Introduction to Written Discourse Analysis*. Routledge, London.

Hunston, Susan and Gill Francis. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. John Benjamins, Amsterdam.

Johnson, Christopher. 2001. Review of Susan Hunston and Gill Francis, *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. *Computational Linguistics*, 27(2):318–320.

Johnson, Jeff. 2000. *GUI Bloopers: Don'ts and Do's for Software Developers and Web*

Designers. Morgan Kaufmann, San Francisco.
Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.
Maedche, Alexander and Steffan Staab. 2000. Semi-automatic engineering of ontologies from text. In *Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering (SEKE 2000)*, Chicago, pages 231–239.
Mann, William C. and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.
Marcu, Daniel. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge.
Morin, Emmanuel and Christian Jacquemin. 1999. Projecting corpus-based semantic links on a thesaurus. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 389–396, College Park, MD.
Morris, Jane and Graeme Hirst. 1991. Lexical cohesion, the thesaurus, and the structure of text. *Computational Linguistics*, 17(1):21–48.
Sinclair, John M, editor-in-chief. 1987. *Collins COBUILD English Language Dictionary*. Collins, London.
Sinclair, John M., Michael Hoey, and Gwyneth Fox. 1993. *Techniques of Description: Spoken and Written Discourse: A Festschrift for Malcolm Coulthard*. Routledge, London.
Thelen, Michael and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, pages 214–221.
Winter, Eugene. 1982. *Towards a Contextual Grammar of English: The Clause and Its Place in the Definition of Sentence*. Allen and Unwin, London.

Graeme Hirst is book review editor of *Computational Linguistics*. His research topics include the problem of near synonymy in lexical choice and the use of lexical relations in intelligent spelling correction. Hirst's address is Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 3G4; e-mail: gh@cs.toronto.edu; URL: http://www.cs.toronto.edu/∼gh.