Book Reviews

Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins

Marie-Hélène Corréard (editor)

Grenoble, France: EURALEX, 2002,
viii+247 pp; paperbound, ISBN
2-9518583-0-2, €30.00
(www.ims.uni-stuttgart.de/euralex)

Reviewed by
Woody Haynes and Martha Evens
Illinois Institute of Technology

This volume is a festschrift for Sue (B. T. S.) Atkins, who is perhaps best known for her work as general editor of the *Collins-Robert English/French Dictionaries* and a consultant-advisor for the *Oxford-Hachette English/French Dictionary,* which led to a revolution in the construction of bilingual dictionaries. She has also done a great deal to bridge the gap between professional lexicography, academic linguistics, and computational linguistics. Most recently she has been working with Charles Fillmore to build FrameNet and to adapt his ideas for use in a dictionary framework. The lead-off paper in this volume is Atkins's keynote address from the 1996 EURALEX meeting, the inspiration for this volume. The contributions of the other authors, all of them new papers, examine how lexicography is responding to Atkins's call for a "radical new type of dictionary" in her 1996 address. They consider the current state of dictionaries, discuss their strengths and weaknesses, and describe new computational tools that are facilitating exploration in new directions and providing new insights.

The increasing availability of diverse electronic text has had a profound effect on the process of creating dictionaries, both in the compilation process and in the depth and structure of the result. It is true that most of the papers in this volume talk about how to use natural language processing in lexicography rather than how to make use of various lexical resources in natural language processing. But we believe that the volume includes much useful material for all those whose work in computational linguistics makes them consumers of lexical resources, since it discusses many of the current ideas about how those resources should be constructed. Even those papers that focus on bilingual resources provide many interesting ideas about the practice of lexicography today. Most of the recommendations for bilingual dictionaries can equally address issues involved in the use of machine-readable dictionaries of all kinds and apply to deficiencies seen in all available lexicons.

Atkins's keynote address, entitled "Bilingual Dictionaries: Past, Present and Future," looks at the various types of information available and needed in various types of bilingual and monolingual dictionaries, categorizes it, and argues for a truly electronic dictionary that can adapt itself to the needs of the "multifarious users." Atkins identifies the strengths of current dictionaries in their wealth of information, scholarly work, and concern for the needs of the dictionary user. She sees weaknesses in the redundancy, coverage gaps, inflexible equivalence and collocational selection, distortion caused by disparate needs of source and target languages and by monolingual infor-

mation omitted from bilingual dictionaries, inextensibility of bilingual dictionaries to multilingual dictionaries, lack of integrated thesaural functions, and the user learning curve for dictionary metalanguage.

In "Use and Usability of Dictionaries: Common Sense and Context Sensibility?" Krista Varantola discusses the disparate needs of lay dictionary users and language professionals. She suggests adapting frame semantics, as proposed by Fillmore and Atkins (1998), to facilitate tailoring the electronic dictionary to give users what they need in terms they understand, relying less on context-free, impenetrable text definitions.

Alain Duval, in "La métalangue, un mal nécessaire du dictionnaire actif," the only paper in the volume not in English, addresses the problems of communicating with the user of a bilingual dictionary. The new bilingual dictionary tries to function as an "active dictionary" that supports the user who is trying to generate text in a second language, while still doing the job of the "passive dictionary" that helps the user who is merely trying to understand that language. Duval illustrates the differences between the old and new with examples from several older bilingual dictionaries and points out the advantages of the new approach for users as well as the demands on the user who must understand the expanded metalanguage.

In "Word Groups in Bilingual Dictionaries: OHFD and After," Richard Wakely and Henri Béjoint describe their approach to usage notes in the *Oxford-Hachette French Dictionary.* They discuss their method of identifying lexical sets exhibiting sufficient size, frequency, and behavior commonality. These sets could then be described once with the entries for each set member pointing to the page containing the usage note. They note the pluses and minuses of this approach in terms of practicality, convenience, and usability.

In "Examples and Collocations in the French 'Dictionnaire de langue,'" A. P. Cowie, current editor of the *International Journal of Lexicography,* looks at the treatment of examples in a number of French monolingual dictionaries, including *Dictionnaire du français contemporain, Le Petit Robert, Le Grand Robert,* and *Le Trésor.* He contrasts their methods of blending examples constructed by lexicographers with quotations, exact or adapted. He contends that "the richness, diversity and fitness for purpose of examples in *Le Grand Robert* and *Le Trésor,* especially, are among the finest achievements in modern lexicography."

Juri Apresjan has been a leading figure in lexicography in the Soviet Union and Russia for over 30 years, since he worked with Igor Mel'čuk on the development of the *Explanatory-Combinatory Dictionary.* More recently he has been head of the major Russian machine translation project. In his paper, "Principles of Systematic Lexicography," he argues for the importance of building a systematic lexicon that can interact effectively with a system of grammar rules in the ECD tradition, and he sketches a linguistic basis for this effort.

Charles Fillmore, the creator of frame semantics and the father of the FrameNet lexical resource (Fillmore and Atkins 1998), discusses the problem of "Lexical Isolates," lexical items that "appear to be of unique semantic or syntactic type." He illustrates some of these behaviors with those problem children *let alone, mention, else,* and *ilk.*

In "Sketching Words," Adam Kilgarriff and David Tugwell describe their method of identifying English word sketches from a corpus with part-of-speech tags and a shallow parse, producing an automatic summary of a word's behavior that can assist lexicographers in describing that behavior and can help NLP systems subsequently to perform word sense disambiguation reliably. Each word sketch consists of one of twenty-six word relations, with one, two, or three operands. The salience of a word sketch is defined as a function of mutual information and log frequencies.

In "Good Old-Fashioned Lexicography: Human Judgment and the Limits of Automation," Michael Rundell, editor-in-chief of the *Macmillan English Dictionary*, considers whether the advances in automation of dictionary development will lead to the demise of the lexicographer. He argues against this view with several compelling examples, suggesting that each advance identifies new layers of complexity that depend on the lexicographer for analysis.

Patrick Hanks, lead editor on a number of Collins and Oxford dictionaries, suggests, in "Mapping Meaning onto Use," that frame semantics provides a "richer schema for representing meaning than is used in any current dictionary." He advocates using a syntagmatic organizing principle in adjective and verb dictionary entries "rather than (or rather, in tandem with) perceived meaning."

Gregory Grefenstette is probably best known for his work on cross-language information retrieval and its application to Internet text. Here he presents "The WWW as a Resource for Lexicography" in a wide range of languages and the tools needed to extract lexical information effectively. He argues that it is feasible to port a number of tools such as shallow parsers to other languages, especially those using some variant of the Roman alphabet, and that it is time to get to work on this project, as significant amounts of text begin to appear in a number of previously unrepresented languages.

Most of the papers in the volume view natural language processing as a tool for building lexicons. In "Lexical Knowledge and Natural Language Processing," Thierry Fontenelle talks about what is needed in a lexical database to support natural language processing and discusses where that knowledge can be found in existing lexical resources, especially collocational dictionaries, thesauri, and semantic networks. This leads naturally to the problems of representing knowledge about verb alternations and other collocations, using the lexical functions of the *Explanatory-Combinatory Dictionary* (Apresjan, Mel'čuk, and Zholkovsky 1970) and Fillmore's frame semantics.

Annie Zaenen, principal scientist and area manager for Multilingual Theory and Technology at the Xerox Research Centre in Grenoble and co-author of several books about lexical-functional grammar and natural language understanding, makes a convincing case for a depressing conclusion in her "Musings about the Impossible Electronic Dictionary." She looks at the complexity and pressures stifling progress in the creation of multifunctional lexicons and concludes that current trends will continue to produce disparate resources for disparate consumption rather than a unified lexical database.

This book is a EURALEX production in every way, and it is certainly a success. Anyone interested in lexicography should read this volume. It might have been even better, however, if the editors had given some of Sue Atkins's many admirers on other continents a chance to join in. The occasional typographical error should certainly be overlooked in view of the bargain price, which should allow many readers to buy copies of their own.

References

Apresjan, Juri D., Igor Mel'čuk, and Alexander Zholkovsky. 1970. Semantics and lexicography: Towards a new type of unilingual dictionary. In Ferenc Kiefer, editor, *Studies in Syntax and Semantics*, Reidel, Dordrecht, pages 1–33.

Fillmore, Charles J. and B. T. S. Atkins. 1998. FrameNet and lexicographic relevance. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, pages 417–423.

Woody Haynes is working on problems of word-sense discrimination, which led to his participation in the latest SENSEVAL. *Martha Evens,* his former thesis advisor and a former president of the Association for Computational Linguistics, is the author of a book on lexicography and the editor of the Cambridge University Press book *Relational Models of the Lexicon.* Their address is Department of Computer Science, Illinois Institute of Technology, 10 West 31st Street, Chicago, IL 60616; e-mail: skhii@mindspring.com, evens@iit.edu.

Multimodality in Language and Speech Systems

Björn Granström, David House, and Inger Karlsson (editors)
(Royal Institute of Technology, Stockholm)

Dordrecht: Kluwer Academic
Publishers (Text, speech and language
technology series, edited by Nancy Ide
and Jean Véronis, volume 19), 2002,
ix+241 pp; hardbound, ISBN
0-4020-0635-7, $82.00, £56.00, €89.00

Reviewed by
Michael Johnston
AT&T Labs—Research

Multimodality in Language and Speech Systems is a collection of papers that stem from a summer school on the topic held at the KTH Royal Institute of Technology in Stockholm, Sweden, in July 1999, under the auspices of the European Language and Speech Network (ELSNET). The volume's chapters address a range of related topics, including taxonomies and descriptive frameworks for analysis and examination of multimodal communication (Allwood, Bernsen), experimental analysis of the relationship between speech and hand gesture (McNeill et al.), audio/visual speech perception (Massaro), multimodality in assistive technology (Edwards), descriptions of implemented systems and architectures that support face-to-face multimodal interaction (Thórisson, Granström et al.), and an intelligent workspace (Brøndsted et al.).

As you might expect, given their origin as summer school presentations, the contributions here primarily do not present new work but rather summarize the authors' research programs or overviews of subareas of the field. As such, the volume is a good introduction to an increasingly important area of speech and language research and provides a solid entry point for more detailed reading. It should be of interest both for use in teaching and for researchers and scholars seeking an introduction to this area.

It should be noted that the contributions in this volume focus primarily on face-to-face multimodal interaction and do not provide an overview of other areas of multimodal interaction such as pen or voice interfaces to mobile devices. Also, the volume does not provide a detailed overview of computational models of multimodal language understanding and multimodal output generation. André (2003) provides an overview of these areas and could be used in teaching along with this volume, readings from Maybury and Wahlster (1998) and Cassell et al. (2000) to provide a more complete overview of the issues, theory, and practice of multimodal systems.

The chapter by Allwood, "Bodily Communication—Dimensions of Expression and Content," illustrates how body movements are essential in interactive face-to-face communication and argues for going beyond analysis of signaled, discrete, written symbols to develop a fuller picture of human communication. The article provides an excellent overview of research on bodily communication over the last century and presents a descriptive framework for analysis of multimodal communication. This framework combines Peirce's division of indexical, iconic, and symbolic information with dimensions of intentionality and awareness (indicate/display/signal). Allwood's contribution clearly illustrates the complexity of the "simultaneous multidimensional coupling" between multiple media of expression and multiple levels of content in face-to-face communication. This point is highly relevant for computational work, since it

explains why embodied conversational systems are so challenging to build: Failure to capture this complexity will lead to unnatural and stilted behavior on the part of artificial-agent communicators.

Like Allwood's, the chapter by Bernsen, "Multimodality in Language and Speech Systems—From Theory to Design Support Tool," provides a framework that can be used in the analysis of multimodal communication and the design of multimodal interactive systems. Whereas Allwood addresses the complexity of face-to-face communication, Bernsen addresses the broader range of interaction between humans, other humans, and machines, including graphical presentations and haptics. The goal of Bernsen's research program is to determine the basic properties of input and output modalities and from these to derive a comprehensive, relevant, and intuitive taxonomy of modalities and modality combinations (modality theory) and to use this theory to aid interaction designers in selecting which representational modalities to use for a given task, context, and user. This chapter provides a highly detailed elucidation and exemplification of a theory and taxonomy of output modalities and briefly describes how this has been used in the development of a hypertext encyclopedic reference tool to aid interaction designers. A number of asymmetries between output modalities and input modalities are addressed but, unlike for output, a comprehensive theory and taxonomy of multimodal input is not yet available. The chapter also summarizes research (Bernsen 1997; Bernsen and Dybkjær 1999) that shows how modality theory accounts for the great majority of claims made in the literature regarding speech functionality. One interesting aspect of Bernsen's modality theory is that, given the top-down development of the taxonomy from theoretical principles, it enables not just analysis of commonplace modalities, but also exploration of new kinds of modalities and modality combinations.

The chapter by McNeill et al., "Dynamic Imagery in Speech and Gesture," argues that human hand gestures are part of our thinking process and that speech and gesture are 'co-expressive': deriving from the same semantic source but able to express different aspects of it. This position is supported by results using the experimental paradigm developed by McNeill, Quek, and colleagues, which combines video-based motion tracking techniques with psycholinguistic analysis of discourse. The chapter presents the experimental method and analysis in detail but provides less detail on the underlying psycholinguistic theory. For this, the reader might want to consult other works (McNeill 1992, 2000). The experimental analysis demonstrates how hand use correlates tightly with the semantic content of discourse. In particular the kind of synchrony (antisymmetry or mirror symmetry) is shown to provide cues for discourse segmentation. Principles are also developed for analysis of the gesture signal, including a 'dominant motion rule' used to determine whether small hand movements are significant.

The chapter by Massaro, "Multimodal Speech Perception: A Paradigm for Speech Science," presents a very clear overview of work on audio/visual speech perception by Massaro and colleagues. The central tenet of the approach is that when evidence from multiple modes, such as audible and visible speech, are combined, the influence of one modality is greater to the extent that the other is ambiguous or neutral. This is captured by a formal model, the fuzzy logic model of perception (FLMP). The core of the chapter is the presentation of the results of a series of experiments that validate the FLMP as an accurate description of multimodal perception. The experiments address the combination of audible speech with lip movement, integration of written text and speech, word recognition, combination of paralinguistic and linguistic cues, and the combination of auditory and facial cues in the perception of emotion. The McGurk effect is also addressed. This chapter provides an excellent introduction to

the program of research pursued by Massaro and colleagues over the last 20 years and provides an entry point for more detailed reading in various books and articles such as Massaro (1998).

Edwards's chapter, "Multimodal Interaction and People with Disabilities," provides a clear (and inspiring) overview of the ways multimodal interface technology has been or could be applied to assisting users with sensory disabilities. The chapter starts with a clear presentation of the properties of different sensory channels and their relationship to modalities of communication and goes on to present a series of examples of interfaces that map one mode into another or use a combination of modes in order to assist people with disabilities.

The chapter by Thórisson, "Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action," addresses the complex problem of modeling turn-taking behavior in multimodal dialog. Starting from literature on human-human interaction, a series of hypotheses are developed regarding the properties of turn-taking behavior. The turn-taking mechanism is characterized as anticipatory, multi-level, highly parallel, and opportunistic. It involves logical combination of multiple sensory features and cues and receives higher (temporal) priority than content analysis and interpretation. Thórisson goes on to show how these hypotheses can be captured in a computational model in which interaction processing is split into three cooperating layers (reactive, process control, content) with differing temporal priorities, and he describes the implementation of the model in the Gandalf prototype. This is an interactive guide to the solar system that supports face-to-face multimodal communication with a synthetic character. A great deal of detail on the implementation is provided, though it is quite densely packed, so the reader may also want to consult Thórisson (1996; 1999) for a fuller understanding of the approach.

The chapter by Granström et al., "Speech and Gestures for Talking Faces in Conversational Dialogue Systems," provides a concise overview of work on audio-visual speech synthesis at KTH. Like Cohen and Massaro (1993), Granström et al.'s approach employs direct parameterization of a graphical model of the face (Parke 1982). In addition to presenting their approach to facial animation and audio-visual synthesis, the authors summarize two perceptual experiments. The first experiment (Teleface) examines the role of visual synthesis in speech intelligibility and its use as an aid to hearing-impaired individuals. For hearing-impaired subjects, adding a synthetic face in addition to the audio channel was found to be almost as much help as adding the natural face. The second experiment explores the relationship between eyebrow movement and intonational phrasing and prominence. Eyebrow movement was found to serve as an independent cue to prominence. The chapter concludes with a description of five different experimental dialogue systems that employ the KTH audio-visual synthesizer (Waxholm, Olga, August, AdApt, and a language tutor) and demonstrates the applicability of the technology to a broad range of application domains.

The chapter by Brødnsted et al., "Developing Intelligent Multimedia Applications," describes a platform for building applications that combine speech and vision developed at the University of Aalborg in Denmark. A sample application for providing campus information is presented. The system supports speech input and output, visual input (camera), and visual output (laser pointer). The authors provide an overview of the underlying system architecture, with a brief description of each component and an interesting example of one type of multimodal application. However, this is primarily a system overview and offers little detail on the approach to multimodal language processing and dialog management adopted.

References

André, E. 2003. Natural language in multimedia/multimodal systems. In Ruslan Mitkov, editor, *Handbook of Computational Linguistics*. Oxford University Press.

Bernsen, Niels Ole. 1997. Towards a tool for predicting speech functionality. *Speech Communication*, 23:181–210.

Bernsen, Niels Ole and Laila Dybkjær. 1999. Working paper on speech functionality. Technical report, Esprit Long-Term Research Project DISC Year 2 Deliverable D2.10, University of Southern Denmark.

Cassell, Justine, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill. 2000. *Embodied Conversational Agents*. MIT Press, Cambridge.

Cohen, Michael M. and Dominic W. Massaro. 1993. Modeling co-articulation in synthetic visual speech. In Nadia Magnenat-Thalmann and Daniel Thalmann, editors, *Models and Techniques in Computer Animation*. Springer-Verlag, Tokyo.

Massaro, Dominic W. 1998. *Perceiving Talking Faces: From Speech Perception to Behavioral Principle*. MIT Press, Cambridge.

Maybury, Mark and Wolfgang Wahlster. 1998. *Readings in Intelligent User Interfaces*. Morgan Kaufmann, Los Altos, California.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago.

McNeill, David. 2000. Growth points, catchments, and contexts. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, 7(1):22–36.

Parke, Frederic I. 1982. Parameterized models for facial animation. *IEEE Computer Graphics*, 2(9):61–68.

Thórisson, Kristinn. 1996. *Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge.

Thórisson, Kristinn. 1999. A mind model for multimodal communicative creatures and humanoids. *International Journal of Applied Artificial Intelligence*, 13(4–5):449–486.

Michael Johnston is principal technical staff at AT&T Labs—Research. His research over the last seven years has focused on the theoretical basis and implementation of multimodal systems, and he leads the MATCH multimodal interface project. Johnston's address is Room E101, AT&T Labs—Research, 180 Park Avenue, Florham Park, NJ 07030, USA; e-mail: johnston@research.att.com.

The Semantics of Relationships: An Interdisciplinary Perspective

Rebecca Green, Carol A. Bean, and Sung Hyon Myaeng (editors)
(University of Maryland, National Library of Medicine, and Chungnam National University)

Dordrecht: Kluwer Academic Publishers (Information science and knowledge management series, edited by J. Mackenzie Owen, volume 3), 2002, xviii+223 pp; hardbound, ISBN 0-4020-0568-7, $97.00, £66.00, €105.00

Reviewed by
Maria Lapata
University of Edinburgh

This book is an edited selection of papers presented at the ACM SIGIR Workshop entitled "Beyond Word Relations," held in Philadelphia in July 1997. Two books arose from this workshop. The first volume, *Relationships in the Organization of Knowledge* (Bean and Green 2001), placed emphasis on thesaural relationships and their role in knowledge organization theory and practice. The second volume, reviewed here, offers an interdisciplinary perspective on relationships and discusses theoretical as well as practical issues concerning their inventory, organization, semantics, and use in real-world applications.

The book consists of 12 chapters organized in three parts, with each part discussing relationships from a different angle. The first part (chapters 1–4) concentrates on relationships and their types. Chapter 1 (Cruse) discusses hyponymy (the inclusion of one semantic class in another), perhaps the most fundamental relation in the organization and representation of meaning. Here we not only learn that the relationship is nearly ubiquitous in human conceptual structures of all kinds, but also that despite its predominance, the relationship has resisted complete characterization. Cruse reviews several definitions of hyponymy and asks important questions: What is the nature of the units related by hyponymy? Are they lexical or conceptual? What are the varieties of hyponymy? Chapter 2 (Fellbaum) examines troponymy, a relationship that characterizes verbs; it is a particular kind of entailment in that every troponym X of a verb Y also entails Y. Unlike nouns, verbs do not seem obviously related in terms of the is-a relation but rather in terms of a variety of manner relations (e.g., *walk* differs from *run* along the dimension of speed). Subrelations of troponymy are discussed, together with the troponymic organization of verbs in WordNet (Fellbaum 1998). Chapter 3 (Pribbenow) investigates meronymy (i.e., the "part-of" relation), discusses the role of parts in human cognition, and introduces classical extensional mereology as an established theory of formalizing parts. Chapter 4 (Khoo, Chan, and Niu) presents a broad overview of the cause-effect relation together with an extensive survey of how the relation can be lexicalized in text.

The second part (chapters 5–8), perhaps the least coherent of the three, discusses relations as exemplified in cognitive semantics, their comparison across ontologies, the notions of identity and subsumption, and a logical system for semantic relationships. Chapter 5 (Green) gives an overview of the conceptual units of cognitive semantics: image schemata, basic-level concepts, and frames. Once the conceptual machinery is in place, linguistic phenomena such as metonymy and metaphor are analyzed. Chapter 6

(Hovy) addresses important methodological questions: What are the characteristics that matter when one describes an ontology? How can one compare ontologies or sets of relations to one another? To address the first of these two questions, Hovy proposes a system of features for the description of ontologies that separate form, content, and usage. For the second, he proposes a methodology by which ontologies can be compared. Chapter 7 (Guarino and Welty) revisits the subsumption relationship from an ontological perspective. Rather than focusing on the relation proper, Guarino and Welty discuss the nature of its arguments and establish the conditions under which subsumption is a well-founded relation. The notions of unity, identity, and essence are discussed at length. Chapter 8 (Jouis) presents a logical system for semantic relationships. The semantic system is based on a set of semantic primitives (types, relations, and properties), and relations are characterized in terms of their functional type, algebraic properties, and relations with other entities.

The third part focuses on applications that make use of semantic relations, methods that automatically discover relations from corpora or knowledge bases, and tools for visualizing relations. Chapter 9 (Evens) explores the use of thesaural information for information retrieval. The types of thesauri used in information retrieval applications are comprehensively reviewed (e.g., Roget's thesaurus, Casagrande and Hale's (1967) relational models, WordNet), together with methods for automatically expanding queries and constructing thesauri. Chapter 10 (Khoo and Myaeng) discusses in-depth methods for identifying semantic relations automatically through pattern matching, again for the purposes of information retrieval. Different types of patterns are investigated (linear, graphical, thematic role), and a methodology for constructing these patterns is presented. Chapter 11 (McCray and Bodenreider) describes the Unified Medical Language System (UMLS) knowledge resources. This is a very clear exposition of a complex system that consists of varied, interrelated and heterogeneous knowledge sources. Descriptions of the Metathesaurus, Semantic Network, and SPECIALIST lexicon are given, together with an example of how UMLS can be used to create specialized semantic networks for a single concept (e.g., heart). Finally, Chapter 12 (Hetzler) explores how visualization can aid the exploration and discovery of relations as well as their expansion and understanding. An informative overview of software for visualizing the presence and absence of relationships is presented.

This book does exactly what its title suggests: it investigates the semantics of relations from an interdisciplinary perspective. It addresses a variety of topics ranging from the theoretical and ontological underpinnings of semantic relations to their visualization, it cuts across research communities, and it does a good job of combining introductory and historical material with technical substance. Although the three parts are loosely related under the general theme of relations, they are autonomous and can be read independently. I would have welcomed more technical detail and more emphasis on evaluation in the third part. The first part notably omits a discussion of antonymy. The various chapters in this book are well-structured and succeed in most cases to present succinctly the topic under discussion, its history and its present. I would have liked to see some more discussion on cross-linguistic aspects and the future of semantic relations. What are the exciting new directions and novel applications? What are the current limitations?

The book should be of interest to theoretical linguists, logicians, and philosophers of language and also to computational linguists and computer scientists. The book is accessible and generally well written. Graduate students with an interest in semantics and their applications will find it useful. I can also imagine some of the chapters being included as reading material for courses on information retrieval and extraction or computational semantics. The lack of an index of authors is compensated for by a

relatively thorough index of terms and rich bibliographic references at the end of each chapter.

References

Bean, Carol A. and Rebecca Green, editors. 2001. *Relationships in the Organization of Knowledge*. Dordrecht, Kluwer Academic Publishers.

Casagrande, Joseph B. and Kenneth L. Hale. 1967. Semantic relations in Papago folk-definitions. In Dell H. Hymes and William E. Bittle, editors, *Studies in Southwestern Ethnolinguistics: Meaning and History in the Languages of the American Southwest*. Mouton, The Hague, The Netherlands, pages 165–193.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Maria Lapata is a research fellow at the University of Edinburgh, Division of Informatics. Her research interests include semantic knowledge acquisition, linguistically informed statistical methods for ambiguity resolution, and computational psycholinguistics. Lapata's address is Division of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH3 9LW, U.K.; e-mail: mlap@inf.ed.ac.uk.

Recent Advances in Computational Terminology

Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme (editors)
(Université Toulouse-le-Mirail, CNRS Orsay, and Université de Montréal)

Amsterdam: John Benjamins (Natural language processing series, edited by Ruslan Mitkov, volume 2), 2001, xviii+379 pp; hardbound, ISBN 1-58811-016-8, $99.00

Reviewed by
Robert Gaizauskas
University of Sheffield

This collection of papers derives from the *Proceedings of the First Workshop on Computational Terminology* (Computerm '98), held at COLING-ACL '98 in Montreal, but is a substantial revision thereof. The current volume comprises seventeen papers plus a brief introduction by the editors. The original workshop proceedings also had seventeen papers. However, seven of these original papers have disappeared, and seven new papers have taken their place. Furthermore, of the remaining papers, most have been significantly extended. Thus, this book should not be thought of as a simple reissue, in hardcover, of the workshop proceedings.

The words *Recent Advances* in the title might be taken to suggest brave strides forward in a clear-cut research program. Nothing could be further from the truth. This is an area of largely pretheoretical research, in which researchers are struggling bravely to use computational techniques to gain some foothold in dealing with the protean complexities of real lexical usage in a variety of technical domains and in a variety of applications. Consequently the book reads a bit like the conversation of the proverbial blind men feeling an elephant, each describing the part he is feeling. This is not meant to be a criticism, for probably nothing else is possible at this time, and besides, this tends to be a feature of edited collections. It does mean, however, that a reader should not come to this volume expecting to find a coherent account of the research issues and approaches in computational terminology. There should be something in here for everyone with any interest in terminology; the danger, however, is that there may not be a meal for anyone.

Classifying the work reported in this volume is not easy. I shall cluster the papers along two dimensions, a major dimension—the task or problem addressed—and a minor dimension—the intended application. This crude structuring should help to convey some notion of the scope and content of the work. In order of ascending complexity the problems addressed by papers in the collection can be characterized as (1) term extraction—the problem of extracting a list of all and only the terms from texts in a given domain, (2) synonymy detection or semantic clustering—the problem of recognizing which terms are synonyms or belong to the same semantic class or cluster, and (3) term-oriented knowledge extraction from text—the problem of building knowledge structures in technical domains, identifying the underlying conceptual entities, attributes, and relations via terminology. The principal application areas addressed by the papers are information retrieval, terminology construction and maintenance, machine translation, automatic index extraction, and automatic abstract generation.

Consider first term extraction, the most basic of the three preceding tasks. Automatic term extraction has potential application in automatic indexing, either for

back-of-book indices or for document-collection navigation, and also for compiling controlled vocabulary terminologies such as are used in, for example, medical coding applications. Several papers address this topic. Most generically, a review paper by M. Teresa Cabré Castellví, Rosa Estopà Bagot, and Jordi Vivaldi Palatresi reviews twelve current term extraction systems, including well-known systems such as LEXTER, FASTR, TERMIGHT, and TERMS, giving a brief description of each, as well as a contrastive analysis. A paper by Lee-Feng Chien and Chun-Liang Chen addresses the problem of incremental update of domain-specific Chinese term lexicons from on-line news sources. Terms are identified and allocated in real time to topic-specific lexicons corresponding to news categories, using highly efficient data structures called PAT trees. To be acceptable for a specific lexicon a term must be *complete* (have no left or right context dependency and have an internal association norm above threshold) and must be *significant* (have a relative frequency in a document collection corresponding to the target lexicon that compares favorably to its relative frequency in a general reference collection). Béatrice Daille supplies a linguistically interesting paper on relational adjectives as signals of terms in French scientific text. Contrasting, e.g., *production importante* ('significant production') with *production laitière* ('dairy production'), she argues that in the latter type of construction, such relational adjectives frequently signal terms. She goes on to describe an automatic technique for identifying such terms that is based on looking for paraphrases of the relational adjective + noun expressed as noun + prepositional phrase, where the complement of the preposition is the nominal form of the relational adjective (so, *production du lait*). A paper by Diana Maynard and Sophia Ananiadou extends their earlier work on term recognition by operationalizing two intuitions about the role of context in termhood: first, that a candidate term that has other candidate terms in its local context is more likely to be a term, and second, that a candidate term that is similar in meaning to domain-specific terms in its local context is more likely to be a term. Toru Hisamitsu and Yoshiki Niwa focus on the specific problem of extracting terms from parenthetical expressions in Japanese news wire text. In expressions of the form $A(B)$, B might or might not be an abbreviated form of A. Segmentation problems in Japanese mean that if B is not correctly recognized as an abbreviation it will be oversegmented into single characters, causing real problems for IR systems. Hisamitsu and Niwa propose a neat solution to the problem of identifying which parenthetical expressions are genuine abbreviations that is based on a combination of statistical and rule-based techniques. Finally, a paper by Hiroshi Nakagawa carefully compares two techniques for term extraction, one based on earlier work by Frantzi and Ananiadou and the other an interesting new proposal that assesses termhood according to how productive a noun in a candidate term is in occurring in many other distinct terms.

The second of the three broad problems or tasks introduced above is the problem of synonymy detection or semantic clustering. Here the problem is not just to discover terms in text, but to relate them in basic ways. Clearly this capability is significant for information retrieval, in which documents similar in meaning to a query, but differing in expression, must be retrieved. Such a capability is also relevant, however, for automatic index creation and for automatic abstracting. Again, several papers in the collection address this topic. Akiko Aizawa and Kyo Kageura propose a technique that cleverly exploits parallel Japanese-English keyword pairs associated with academic papers to build multilingual semantically related keyword clusters for use in monolingual or cross-lingual IR applications. Peter Anick proposes to use **lexical dispersion**, a measure of the extent to which a given word is used in multiple NP constructs, to identify generic concepts in retrieval results and to structure these results accordingly. Hongyan Jing and Evelyne Tzoukermann present a stimulating new

approach to a classical problem in IR. Intuition suggests that both collapsing variant morphological forms of words and distinguishing different word senses ought to improve retrieval. But previous attempts to do so, through stemming and sense disambiguation, have not led to a reliable increase in performance. The authors present an approach based on full morphological analysis, rather than stemming, and on using context vectors to represent word sense distinctions and to determine whether identical strings in the query and the document should be matched. The approach shows improvement over a more conventional model in which string identity is the only test of synonymy. Thierry Hamon and Adeline Nazarenko start with an existing lexical resource containing synonym links and a term extractor and use these to bootstrap term synonym sets by (1) analyzing each compound term in a technical corpus into a head + expansion (modifiers) and then (2) forming candidate synonyms of the compound by combining (a) synonyms of the original head with the original expansion, (b) the original head with synonyms of the original expansion, and (c) synonyms of the original head with synonyms of the original expansion. The resulting synonym sets are to be used in a document-consulting system to help users navigate complex technical documents. Adeline Nazarenko, Pierre Zweigenbaum, Benoît Habert, and Jacques Bouaud contribute a paper that describes an approach to classifying unknown words in a medical corpus into one of the eleven top-level semantic categories in the SNOMED hierarchical terminology. They parse the corpus for NPs, extract dependency relations between the words in the parsed NPs, e.g., W_1 R W_2, and construct a graph wherein words are the nodes and edges are labeled with shared contexts between the connected words: W_1 and W_3 share a context if for some R and W_2, both W_1 R W_2 and W_3 R W_2 are attested in the corpus. In this similarity graph, words whose semantic category is known from the SNOMED resource are labeled with their category, and categories are then propagated to uncategorized nodes via a voting mechanism between the uncategorized nodes' nearest neighbors. Finally, Michael Oakes and Chris Paice describe a technique for validating terms that can occur in particular semantic roles, or slots, in an information extraction–like template structure designed to capture details of scientific papers for use in generating abstracts. Starting with an initial, corpus-derived thesaurus containing domain-specific high-frequency words and multiword units (MWUs), each manually tagged with its semantic role, the MWUs are analyzed to reveal any that contain as substrings words or shorter MWUs already in the thesaurus. For such MWUs a semantic grammar rule is generated whose pattern is the MWU with the substring replaced by its semantic role and whose action is to label matching strings with the semantic role of the MWU. Such rules, which implicitly define a class of semantically equivalent terms, generalize the thesaurus beyond observed examples and are used to validate proposed slot fillers in the template.

The third problem area, and the most challenging, is that of building knowledge-rich terminologies—terminologies that contain not only terms, but attributes and relationships of the concepts denoted by the terms, frequently for use in applications requiring controlled terminologies. James Cimino contributes a paper describing the methodology employed in maintaining a large-scale knowledge-based controlled medical terminology used to encode patient data and to provide aggregation classes for a variety of applications, such as billing and decision support. In such a critical and knowledge-rich environment, terms cannot be automatically added to the terminology as a consequence of language processing. However, Cimino describes how simple language processing, together with knowledge-based reasoning, can be used to guide a terminologist in the process of, for example, adding the name of a new drug. Anne Condamines and Josette Rebeyrolle describe a corpus-driven approach to constructing a terminological knowledge base. First they use Bourigault's LEXTER to identify

candidate terms, then initiate a search for conceptual relationships among them. Taxonomies are constructed by using a fixed set of linguistic patterns to identify candidate hypernymic and meronymic binary relations. Then pairwise comparisons are made between terms in different taxonomies, and recurrent contexts in the corpus are sought in which these term pairs co-occur. If such contexts are found, a conceptual relationship is proposed, linguistic patterns are created to match the context, and these patterns applied to the corpus to identify new terms. The process is then repeated until no more relationships or terms are found. Although highly suggestive, this paper is unclear in critical places, particularly regarding which steps are carried out manually and which automatically. Finally, Ingrid Meyer presents a framework for building knowledge-rich terminological dictionaries. Her approach is firmly semiautomatic, tools being provided to assist, rather than replace, a human terminologist. The method depends on acquiring knowledge patterns, which may be lexical, grammatical, or paralinguistic (relying on, e.g., punctuation), to find **knowledge-rich contexts** from which hypernymic or other attribute or relational knowledge may be extracted. Such patterns are acquired through an iterative manual process of refinement in conjunction with a corpus.

Not fitting neatly into the above classification are two papers on bilingual term alignment for machine translation (MT). This is an important application area for terminology systems, as the translation of terminology-laden technical documents is commercial MT's bread and butter and an area in which human translators' lack of domain-specific knowledge is likely to be a bottleneck. Eric Gaussier's paper gives a general overview of issues faced in bilingual terminology extraction from a parallel corpus, particularly choices between (1) extracting terms in each language independently, then aligning terms, or (2) parsing terms in one language, then projecting, by alignment, terms onto the second language, or (3) parallel parsing. He explores an idea, referred to as **pattern affinities**, that candidate terms expressed via one syntactic pattern in one language are more likely to be rendered in the other language by some other specific syntactic pattern but shows via an implementation of this idea using the EM algorithm that results are not significantly improved. David Hull, in a very clear and convincing piece, describes a method of bilingual lexicon construction from translated sentence pairs that relies on term extraction in the source language and a probabilistic word translation model to propose term translations in the target language. Although not perfect, this model can, the author argues, lead to significant productivity gain in constructing bilingual term lexica when used in a semiautomated mode by a human terminologist.

This is a wide-ranging collection, and the editors are to be congratulated for pulling together so much interesting material. That said, there a few reproaches to be leveled at them too. First, the level of copyediting is very poor. Spelling mistakes and minor grammatical errors abound; figures and tables have incorrect captions; formulas have undefined terms. At best this is irritating; at worst it seriously impedes understanding and conveys an impression of sloppiness that undermines the reader's trust. Perhaps this falls in the crack between what the editors and the publishers feel is their responsibility. But one expects somewhat better. Second, the book would have been much more readable and more generally useful had the papers been structured into related subareas, with introductory overviews in each area setting the stage for, and comparing and contrasting, the relevant papers. As it is, the editors have opted to present the papers in alphabetical order by first author's surname, hardly the most cognitively compelling of structural principles. The editors' introduction is a step in the right direction, but a small one, and given the wide range of topics addressed, some further analysis and guidance would have been welcome. As a consequence, although this is

a book I would regret not having in my university library, it is not one I would regret not owning myself.

Robert Gaizauskas is Professor of Computer Science at the University of Sheffield. His research interests are in applied natural language processing, specifically information extraction, most recently concentrating on biomedical texts. Gaizauskas's address is Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello St., Sheffield, S1 4DP, U.K.; e-mail: R.Gaizauskas@dcs.shef.ac.uk.