# Book Reviews

## Early Years in Machine Translation: Memoirs and Biographies of Pioneers

**W. John Hutchins (editor)**

Amsterdam: John Benjamins (Studies in
the history of the language sciences,
edited by E. F. Konrad Koerner, volume
97), 2000, xii+400 pp; hardbound, ISBN
1-58811-013-3 and 1-55619-013-3, $95.00

*Reviewed by*
*Warren J. Plath*
*IBM T. J. Watson Research Center*

In the preface to this extensive collection of memoirs and biographies, the editor describes its purpose as follows:

> The aim when compiling this volume has been to hear from those who participated directly in the earliest years of mechanical translation, or 'machine translation' (MT) as it is now commonly known, and, in the case of those major figures already deceased, to obtain memories and assessments from people who knew them well. Naturally, it has not been possible to cover every one of the pioneers of machine translation, but the principal researchers of the United States, the Soviet Union, and Europe (East and West) are represented here. (page vii)

The collection includes contributions by some 26 individuals who were involved in MT in the 1950s and 1960s, augmented by an introduction and articles by the editor, John Hutchins, on Warren Weaver, Yehoshua Bar-Hillel, and Gilbert King. Along with accounts of the origins and histories of their respective research projects, the authors have provided numerous personal details and anecdotes as well as a number of photographs, contributing significantly to the richness of the overall presentation.

In his introduction, "The First Decades of Machine Translation: Overview, Chronology, Sources," Hutchins begins by noting the seminal significance of early MT work for computational linguistics, natural language processing, and other areas, as well as the wide variety of backgrounds, aims, and approaches of the pioneers. The overview section contains a brief account of major features of and influences on MT work of the period, including theoretical frameworks, technological constraints, funding sources, and evolving goals. This is followed by the chronology section—a compact history of MT from its beginnings to the mid-1970s—and the sources section, consisting of three pages of bibliographic references. Taken as a whole, the introduction provides the reader with valuable background material that is conducive to a fuller appreciation of the articles that follow.

The articles are grouped geographically, beginning with U.S. pioneers and proceeding to those from the Soviet Union, the United Kingdom, Western and Eastern Europe, and Japan. The U.S. group is further partitioned, roughly chronologically, into three subgroups: first, the earliest pioneers (Warren Weaver, Erwin Reifler, Victor Yngve, and Anthony Oettinger); next, individuals with a connection to the Georgetown project (Leon Dostert, Paul Garvin, Michael Zarechnak, Tony Brown, and Peter Toma); and then a set of researchers who became active somewhat later (Winfred Lehmann, David Hays, Gilbert King, and Sydney Lamb).

The first article, "Warren Weaver and the Launching of MT: Brief Biographical Note," by the editor, is a biographical sketch focusing on Weaver's famous 1949 memorandum and its catalytic impact in launching the field. The second contribution, "Erwin Reifler and Machine Translation at the University of Washington," was written by Reifler's former colleague Lew R. Micklesen, a Slavic linguist. The first half of the article focuses on Reifler—his background in Europe and China, his early enthusiasm for MT, and his work on German and Chinese—while the remaining portion deals primarily with the author's own experiences in developing the original version of the Russian-English dictionary for the Rome Air Development Center, initially at the University of Washington under Reifler and later at IBM under Gilbert King.

"Early Research at M.I.T.: In Search of Adequate Theory," by Victor H. Yngve, provides a detailed account of the author's wide-ranging activities in the field of MT during its early years. The narrative highlights such contributions as his experiments on gap analysis and random generation, the development of the COMIT programming language, the cofounding and editing of the journal *MT*, and the formulation of the depth hypothesis. The author describes at some length Chomsky's outright rejection of the depth hypothesis and presents a vigorous countercritique of Chomsky's work and of abstract linguistics generally, labeling such approaches "unscientific" (page 68). He concludes the article by advocating what he calls "the new foundations of general linguistics" (page 69), which he has put forth in a textbook (Yngve 1996).

In "Machine Translation at Harvard," the last of the articles on the earliest U.S. pioneers, Anthony Oettinger recounts the history of the Harvard project, including his design and development of the Harvard Automatic Dictionary and the subsequent theoretical and applied work in the area of syntax. The article also includes interesting accounts of his personal experiences, especially in connection with his 1958 visit to the Soviet Union and his later participation as a junior member of the Automatic Language Processing Advisory Committee (ALPAC) of the National Academy of Sciences, whose 1966 report had as great an influence on the course of MT as did Weaver's 1949 memo.

The five articles relating to the Georgetown project are of interest both for their individual content and as a source of sometimes sharply divergent views of events and relationships. "The Georgetown Project and Leon Dostert: Recollections of a Young Assistant," by Muriel Vasconcellos, provides many colorful details on the career and personality of Dostert, the project director, as well as an account of the project's history and its organization into subgroups. The impression conveyed by Vasconcellos is of a well-structured and relatively smoothly functioning operation in which intergroup discussions only occasionally "got rather heated" (page 93). A less idyllic picture emerges from the next article, "Is FAHQ(M)T Possible? Memories of Paul L. Garvin and Other MT Colleagues," by Christine Montgomery, who describes the project as evolving into "a set of armed camps" (page 100) in which the weekly intergroup seminars were "characterized by a lack of harmony . . . overlaid with a veil of secrecy and distrust" (page 102). The main focus of the article, however, is on Paul Garvin and what the author views as the present-day legacy of his empirically oriented approach to machine translation.

In "The Early Days of GAT-SLC," Michael Zarechnak describes the origins of the main Georgetown translation system, which he and his team developed for Russian-English MT. It consisted of GAT (general analysis technique), the linguistic component, and SLC (simulated linguistic computer), the computational component developed by A. F. R. Brown. The article includes a detailed description of the linguistic approach, illustrated with numerous examples in transliterated Russian. The following contribution, "Machine Translation: Just a Question of Finding the Right Programming Language?" by Antony F. R. Brown, provides an account of the author's development of SLC, along with a sketch of his subsequent career. The final article in the group, "From SERNA to SYSTRAN," by Peter Toma, describes the author's somewhat turbulent career at Georgetown, followed by his subsequent rise to fame and fortune as the developer of SYSTRAN. Although some readers may be put off by the self-congratulatory tone of the presentation, it is nonetheless a compelling story of how an able and highly ambitious individual achieved MT's first commercial success.

The main body of "My Early Years in Machine Translation," by Winfred Lehmann, is an account of the history and research approach of the University of Texas project, which the author founded and led for many years. For this reviewer, however, the two most fascinating sections are "Previous Background" at the beginning (pages 147–149) and "Suspension of Research as a Result of the ALPAC Report" at the end (pages 160–162). The former describes the author's post–Pearl Harbor experiences in the Army translation program as it scrambled to catch up with a huge backlog of intercepted Japanese military and diplomatic messages—a situation eerily parallel to the current government's position vis-à-vis Arabic and Central Asian languages some 60 years later. The latter section contains a rather bitter denunciation of the ALPAC report (National Academy of Sciences 1966), including the remarkable assertion that none of the members of the committee, which included David Hays and Anthony Oettinger, "were prominent in the field" (page 161).

In contrast to the preceding article, "David G. Hays," by Martin Kay, is an enthusiastic summary of Hays's contributions to MT and computational linguistics, including his role in founding AMTCL (Association for Machine Translation and Computational Linguistics), ICCL (International Committee on Computational Linguistics), and the biennial COLING (Computational Linguistics) conferences. This is followed by "Gilbert W. King and the IBM-USAF Translator," by John Hutchins, and "Translation and the Structure of Language," by Sydney M. Lamb, the final two articles on American MT pioneers. Hutchins provides a brief account of King's oversimplified approach to translation, with its minimal linguistics and reliance on special-purpose hardware; Lamb describes the Berkeley translation project, emphasizing its lexical organization techniques and his evolving view of language as a network of relationships.

The five contributions by MT researchers from the former Soviet Union provide an interesting and diverse set of perspectives both on the technical approaches and achievements of their respective groups and on the political conditions under which they operated. The authors of the first three articles (Olga Kulagina, Igor Mel'čuk, and Tat'jana Mološnaja) were all associated with Ljapunov's group at the Institute of Applied Mathematics in Moscow, which began work on French-Russian and English-Russian MT in the mid-1950s. Raimund Piotrovskij, the author of the fourth article, was a member of Nikolaj Andreev's group at Leningrad State University, known for its emphasis on development of an intermediate language to facilitate translation. The final article in the group is by Jurij Marčuk, a former KGB officer who worked on English-Russian machine translation.

In "Pioneering MT in the Soviet Union," Kulagina describes the first-generation French-Russian system FR-I and its dependency tree–based successor FR-II against the backdrop of the rise and subsequent decline of Soviet activity in MT. She attributes the latter trend to a combination of ineffective state support and disenchantment due to the intrinsic difficulty of the problem, rather than to the impact of the ALPAC report.

This assessment stands in marked contrast to those of Mel′čuk and Piotrovskij, both of whom assert that the report led to termination of funding for many MT projects in the Soviet Union. The following article, "Machine Translation and Formal Linguistics in the USSR," by Mel′čuk, is based on the transcript of an extended interview with the editor in 1998. Beyond the value of its technical content, this article is an example of oral history at its best, offering an illuminating and engaging portrait of personalities, relationships, and political conditions as they affected the personal life and professional career of a talented linguist striving to cope with the handicap of his status as a Jew, and later an outright dissident, in the Soviet Union.

The third article in the group, "My Memoirs of MT in the Soviet Union," by Mološnaja, is a very brief piece, notable both for her warm recollections of former colleagues and for a sharp critique of the rival Moscow-based project at the Institute of Precise Mechanics and Computer Technology. In the final two articles, by Piotrovskij and Marčuk, the authors strongly advocate what they consider to be practical approaches to MT, while dismissing much of the work cited in the first three articles as misguided and counterproductive. Thus Piotrovskij, in "MT in the Former USSR and in the Newly Independent States (NIS): Pre-history, Romantic Era, Prosaic Time," criticizes the approach of his former mentor Andreev in Leningrad, as well as that of Mel′čuk and Kulagina, as having "driven us into deadlock" (page 235). Marčuk is even more pointed in his criticism, slipping in an apparent anti-Semitic slur: "In famous traditions of Bolshevism and the Talmud ("he who is not with us is against us") Mel′čuk and his supporters attacked all practical workers in the MT field.... Significantly, after fifty years of MT not a single practical MT system has appeared that uses the 'meaning-text' approach to any significant extent" (page 249).

"The Beginnings of MT," by Andrew D. Booth and Kathleen H. V. Booth, is the first of three articles relating to MT pioneers from the United Kingdom. The account begins with A. D. Booth's early contacts with Warren Weaver in 1946 and 1947 and continues with a brief description of the varied activities of the project that Booth headed at Birkbeck College of the University of London until 1962, enlivened by several anecdotes from that period. The authors go on to describe their administrative and MT research activities at Canadian universities in the 1960s and 1970s, which included large-scale dictionary building.

The next two articles deal with research activities at the Cambridge Language Research Unit (CLRU) dating from the mid-1950s, focusing on the contributions of the botanist R. H. Richens and on those of CLRU's founder and director, the redoubtable Margaret Masterman. In "R. H. Richens: Translation in the NUDE," Karen Sparck Jones reviews and analyzes Richens's key papers, tracing the development of his ideas concerning a semantically based interlingua to their culmination in NUDE: a structured representation conceived of as a semantic net of 'naked ideas'. The author describes how NUDE was used by the CLRU staff in their Italian dictionary and also figured in research in other areas such as thesaurus design. She notes, however, that the group never managed to use it successfully as a vehicle for translation, owing to a failure to deal adequately with syntax and its interaction with semantics, a failure that she largely lays at the doorstep of CLRU's director: "Masterman adopted, however, such an aggressively fundamentalist approach to this whole pattern determination operation, and so resolutely eschewed help from syntax, that she was never able to carry her ideas into effective computational practice" (page 276).

In "Margaret Masterman," Yorick Wilks, although not entirely uncritical, presents a much more favorable picture of Masterman's technical contributions, focusing more on what he views as her seminal ideas than on practical results. He credits her with being some 20 years ahead of her time in advocating such approaches as computational lexicography and parsing by semantic methods, while providing a rather indulgent account of Masterman's more eccentric pursuits, such as her long-term attempts to partition texts on the basis of breath groups and rhetorical figures. The article includes a history of the CLRU, which Wilks considers to be Masterman's principal practical creation and a tribute to her persistence and enthusiasm. Throughout the article, he does an excellent job of bringing this unique character to life, noting at one point that "her ideas were notable . . . for their sheer joyousness" (page 284).

The editor leads off the final segment on "researchers from elsewhere" with "Yehoshua Bar-Hillel: A Philosopher's Contribution to Machine Translation," which chronicles Bar-Hillel's progression from early enthusiast and promoter of MT, through his oft-cited later skepticism, to his ultimately more moderate (and less well-known) views regarding the possibility of high-quality results. Next comes "Silvio Ceccato and the Correlational Grammar," by Ceccato's former disciple Ernst von Glasersfeld. The piece begins with a description of the early attempt of Ceccato's project to construct a Russian-English MT system based on a representation of meaning as a network of operations linked by explicit and implicit connections called correlations. Glasersfeld then goes on to recount his own experiences in the years following the demise of the original project, when he left Italy and attempted to continue Ceccato's approach at the University of Georgia, ultimately using it in experiments with Lana the chimpanzee at the Yerkes Institute in Atlanta.

The next two articles relate to the two major MT projects initiated in France in the early 1960s: first, the short-lived Paris project under Aimé Sestier, and then the decades-long effort led by Bernard Vauquois at Grenoble. "Early MT in France," by Maurice Gross, presents only a very brief sketch of the Paris project, which focused on Russian-French translation and terminated early in 1963 after Sestier became convinced that the task was too difficult to pursue further. In contrast, Christian Boitet's article, "Bernard Vauquois' Contribution to the Theory and Practice of Building MT Systems: A Historical Perspective," provides a relatively detailed picture of both the Grenoble project and the accomplishments of its leader in his various roles as researcher, teacher, MT system builder, and international figure in computational linguistics.

The last three contributions to the collection deal respectively with early MT activities in Czechoslovakia, Bulgaria, and Japan. In "Pioneer Work in Machine Translation in Czechoslovakia," Zdeněk Kirschner recounts the experiences of the MT research group in Prague from the late 1950s into the 1980s as it coped with chronically primitive computing facilities and struggled to survive during the political repression that followed the "Prague Spring" of 1968. The author gives the main credit for the group's accomplishments to Petr Sgall, citing his technical leadership and managerial skills, as well as his personal courage in the face of intense political pressure.

"Alexander Ljudskanov," by Elena Paskaleva, is a highly laudatory account of the personal background and professional career of this Bulgarian pioneer, known more for his theoretical publications and international activities than for his project on Russian-Bulgarian translation. The latter work, in the author's view, might well have come to practical fruition were it not for Ljudskanov's untimely death at the age of 50.

In the final article, "Memoirs of a Survivor," Hiroshi Wada describes the work on English-Japanese MT that he initiated in 1957 at the Electrotechnical Laboratory of Japan. The account covers the varied activities of the project, including the design of computers and optical character recognition (OCR) systems, dictionary building, and translation algorithm development, which culminated a few years later in a capability to translate a range of simple English sentences into Japanese counterparts printed out as strings of kana characters. The article concludes with a brief mention of other MT-related work of that era in Japan, with emphasis on OCR development and kana-kanji translation.

At the end of the book, the editor has included two separate indexes: an index of names, augmented in many instances by birth and (where appropriate) death dates, and an index of subjects. This bipartite organization provides added convenience for the reader who wishes to compare the variety of perspectives on specific persons and events that are offered by the contributors to the collection. This final touch is representative of the thoughtful design and careful editorial workmanship that are characteristic of the volume as a whole. Aside from a very few residual proofreading errors, the only flaw I noticed was the incorrect rendering of the name of the late Asher Opler as "Ashley Opler" (pages xii, 391).

In capturing and preserving this impressively wide-ranging collection of reminiscences, John Hutchins has made a huge and enormously valuable contribution to our understanding of the ideas, personalities, and external forces that shaped the early development of machine translation and computational linguistics and that set in motion many of the activities in those areas that are still ongoing today. I heartily recommend this book not only for readers engaged in those or related fields, but also for anyone with an interest in the history of science.

**References**

Yngve, Victor H. 1996. *From Grammar to Science: New Foundations for General Linguistics*. Amsterdam and Philadelphia: John Benjamins.

National Academy of Sciences. 1966. *Language and Machines: Computers in Translation and Linguistics*. Publication 1416, National Academy of Sciences, Washington, D.C.

*Warren J. Plath* is a research staff member, emeritus, at the IBM T. J. Watson Research Center, where he formerly served as manager of the Theoretical and Computational Linguistics Group. As a graduate student and later faculty member at Harvard in the late 1950s and early 1960s, he was a member of the machine translation project under Anthony G. Oettinger, in which he worked on automatic parsing of Russian. At IBM in the mid-1960s he managed the transition from applied work on Russian-English MT using special-purpose machines to a broad program of research emphasizing linguistic theory, English grammar, and formal parsing systems implemented on general-purpose computers. His subsequent project and individual research included work on natural language question-answering, machine-aided translation, semantic grammars for spoken-language interfaces, and information extraction from news articles on the Web. Plath's address is Mathematical Sciences Department, IBM T. J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598-2103; e-mail: plath@us.ibm.com.

## 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.

## Word Frequency Distributions

**R. Harald Baayen**
(University of Nijmegen)

Dordrecht: Kluwer Academic
Publishers (Text, speech and language
technology series, edited by Nancy Ide
and Jean Véronis, volume 18), 2001,
xxii+333 pp and CD-ROM; hardbound,
ISBN 0-7923-7017-1, $108.00, €117.00,
£74.00; paperwork, ISBN 1-4020-0927-5,
$48.00, €48.00, £30.00

*Reviewed by*
*Geoffrey Sampson*
*University of Sussex*

Baayen's book must surely in the future become the standard point of departure for statistical studies of vocabulary.

Baayen begins with a puzzle that has troubled many investigators who have studied vocabulary richness, for instance, people hoping to find stylistic constants characteristic of individual authors for use in literary or forensic authorship disputes. Naïvely one imagines that the ratio of number of distinct word types in a document to number of word tokens—the "type/token ratio," or as Baayen prefers, exchanging numerator and denominator, the "mean word frequency"—might be a suitable index. It is not, because it is not independent of sample size. In most domains, sample means fluctuate randomly around population means while getting closer to them as sample sizes increase. In natural language vocabulary studies, mean word frequencies systematically increase with sample size even when samples of tens of millions of words are examined.

To make the point concrete, Baayen compares Lewis Carroll's *Alice in Wonderland* and *Alice through the Looking-Glass*. One might hypothesize that greater experience would lead a writer to use a richer vocabulary in a later book, but mean word frequency is actually higher (i.e., type/token ratio lower) in *Through the Looking-Glass* than in *Wonderland*: 10.09 to 10.00. *Through the Looking-Glass*, however, is a somewhat longer book. If just the first 26,505 words are used (this is the length of the earlier book), the direction of the difference in mean word frequencies is reversed: 9.71 to 10.00. Normally, more data give a more accurate picture (of anything); but here the direction of change in frequency, from 9.71 for 26,505 words to 10.09 for 29,028 words, is usual. Can we conclude that Carroll was using a richer vocabulary in the later book, because of the figures for equal-sized samples? Or that he was using a less rich vocabulary, because of the figures for total available samples? Or can we make no inference either way?

A number of scholars have devised formulae more complex than the simple type/token ratio in an attempt to define characteristic constants that are independent of sample size. Gustav Herdan argued in a series of works that were influential in the 1960s that the ratio of the logarithms of number of types and number of tokens was such a constant. Baayen considers "Herdan's law" and various other proposals in the literature, such as G. K. Zipf's, and shows empirically that each is mistaken: All the measures turn out to be dependent on sample size (though one proposed by Honoré [1979] appears to be less so than the others). Conversely, Baayen quotes Naranan and

Balasubrahmanyan (1998, page 38) as claiming that "a word frequency analysis of a text can reveal nothing about its characteristics." Eventually, Baayen is able to show that this negative position is also unjustified; but between that conclusion and the statement of the puzzle lie some two hundred pages of fairly dense mathematics. (This is certainly not a book for the mathematically fainthearted. Baayen does a great deal, though, to help the reader follow him through the thickets. Not only does each chapter end with a summary of its findings, but—unusually for a work that is not a student textbook—Baayen also gives lists of test questions that the diligent reader can work through to consolidate his understanding of the material.)

What lies behind the unusual relationship between type frequencies and sample sizes in the case of vocabulary? Baayen clarifies the situation by an analogy with die-throwing. Think of repeated throws of a single die as a system generating a sequence over the vocabulary "one, two, . . . , six": Baayen plots a graph showing how the expected frequency spectrum (that is, the number of vocabulary elements observed once, the number of vocabulary elements observed twice, . . .) changes as the sequence is extended. For *hapax legomena* (elements of the vocabulary observed once each), the expected figure rises to a maximum of about 2.5 (I am reading approximate figures off Baayen's plot rather than calculating exact figures for myself) at five throws, and then falls back to near zero by 40 throws. For successive elements of the spectrum, the waves are successively lower and later, but the pattern is similar: for *dis legomena* (types observed twice) the maximum is about 1.8 at about 12 throws and close to zero by about 60 throws, and so on. Meanwhile, a plot on the same graph of expected sample vocabulary size rises rapidly and is close to the population vocabulary (i.e., six) by 40 throws. In most domains to which statistical techniques are applied, sample sizes are large enough to involve areas far out to the right of this kind of graph (a serious examination of possible bias in a die would surely involve hundreds of throws), so the special features of its left-hand end are irrelevant. With natural language vocabulary studies, on the other hand, even the largest practical samples leave us in an area analogous to the extreme left-hand end of the die-throwing graph, with numbers of *hapax legomena* (and consequently also *dis legomena, tris legomena*, etc.), as well as vocabulary size, continuing to grow with increased sample size and showing no sign of leveling out.

Using a term borrowed from Khmaladze (1987), Baayen describes achievable sample sizes in vocabulary studies as falling into the "large number of rare events" (LNRE) zone of the sample-size scale. The intuitive meaning of this is fairly clear, and it is made exact through alternative formal definitions. Much of Baayen's book is about the special mathematical techniques relevant to the study of LNRE distributions. (Using these techniques, it turns out that the growth in vocabulary richness between *Alice in Wonderland* and *Alice Through the Looking-Glass*, after truncation to make their length the same, *is* marginally significant.) Not all of the exposition is original with Baayen. One of the many virtues of his book lies in drawing together in one convenient location a clear statement of relevant analyses by others over several decades, often published relatively obscurely. Baayen's chapter 3 presents three families of LNRE models, which are due respectively to J. B. Carroll (1967), H. S. Sichel (1975), and J. K. Orlov and R. Y. Chitashvili (1983a, 1983b). A point that emerges from the book (and that readers of this review may have begun to infer from names cited) is the extent to which, in the late 20th century, this mathematical approach to natural language was a scholarly specialty of the former Soviet Union; in consequence it was largely unknown in the West. There are other channels through which this work has become accessible to the English-speaking world in recent years, notably the *Journal of Quantitative Linguistics*, but that German-based journal, though published in English, has to date attracted limited attention in Britain and North America. The book under review may well be the most significant route by which important Soviet research in our area will become known to English-speaking scholars.

It would be beyond the scope of this review to survey all the issues relating to LNRE distributions that Baayen investigates. For linguists, one particularly interesting area concerns departures from the randomness assumption made by the simpler LNRE models. These pretend, for the sake of mathematical convenience, that texts are constructed by drawing successive words blindly out of an urn containing different numbers of tokens of all possible words in the vocabulary, so that the difficulties to be addressed relate only to the vast size of the urn. Real life is not like that, of course: for instance, from the frequency of the word *the*, the urn model predicts that the sequence *the the* should occur once in every couple of pages or so of text, but in practice that sequence is hardly ever encountered.

If we are primarily interested in overall vocabulary size, one problem that is repeatedly produced by the urn model is that inferences from vocabulary size in observed samples to vocabulary sizes for other, so-far-unobserved sample sizes turn out to be overestimates when samples of the relevant size are examined. Many linguists, particularly after the above discussion of *the the*, will be professionally inclined to assume that this problem stems from ignoring syntactic constraints within sentences, as the urn model does. Baayen demonstrates that this is *not* the source of the problem. If the sentences of *Alice in Wonderland* are permuted into a random order (while preserving the sequence of words within each individual sentence), the overestimation bias disappears. Instead, the problem arises because key words (for *Alice in Wonderland*, some examples are *queen, king, turtle*, and *hatter*) are "underdispersed." Different passages of a document deal with different topics, so topic-sensitive words are not distributed evenly through the text.

The bulk of Baayen's book consists of sophisticated mathematical analysis of the kinds of issues considered in the preceding paragraphs. No doubt what Baayen gives us is not always the last word to be said on some of the questions he takes up, but (as already suggested) it is hard to think that future analyses will not treat Baayen as the standard jumping-off point for further exploration.

Baayen's final chapter (chapter 6) concerns applications, and this is arguably something of an anticlimax. It is natural to want to show that the analysis yields implications for concrete topics, but some of the topics investigated do not seem very interesting other than as illustrations of Baayen's techniques, and some of them apparently lack the LNRE quality that gives the bulk of this book its impact. For natural language–processing applications, probably the most significant topic considered is bigram frequency (Baayen's section 6.4.4), but on this the author has only a very limited amount to add to the existing literature. In terms of general human interest, there is much promise in a section that studies the statistical pattern of references in recent newspapers to earlier years from the 13th century onward and finds a striking discontinuity about the year 1935 "suggesting that this is a pivotal period for present-day historical consciousness." But in the first place, this seems disconnected from the body of the book, because the relevant distributions are not LNRE. Furthermore, the only newspaper identified by name is the *Frankfurter Allgemeine Zeitung*, and although we are told that other newspapers show the same pattern, we are not told which newspapers these are. Finding that Germans perceive a unique historical discontinuity in the 1930s might be a very different thing from finding that Europeans, or Westerners in general, do so.

Nevertheless, this last chapter does also contain important findings that relate more closely to the central concerns of the book. In this chapter Baayen illustrates the sophisticated statistical calculations that he uses in place of naïve type/token ratios, in the quest for characteristic constants of lexical usage. For each of a range of literary works, the calculations yield a curve occupying some portion of a two-dimensional "authorial space" (my phrase rather than Baayen's). With many pairs of separate works by the same author, the resulting curves are satisfactorily close to one another and well separated from curves for other authors: This is true when authors are as different as Henry James (*Confidence* and *The Europeans*) and St. Luke (St. Luke's Gospel and *Acts of the Apostles*). But there are exceptions: H. G. Wells is a case showing that "intra-author variability may be greater than inter-author variability," since the curves for his *War of the Worlds* and *The Invisible Man* are somewhat far apart, and the curves for Jack London's *Sea Wolf* and *The Call of the Wild* are superimposed on one another in the space between the two Wells curves.

The final chapter also contains a number of misprints, which are not self-correcting and may be worth listing here. In a discussion of word length distribution, there are repeated confusions between length 4, length 5, and length 6, on pages 196, 197 (Figure 6.1), 198 (Figure 6.2), and 199; some of the passages indicated may be correct as printed, but they cannot all be correct. On page 204, in a list of Dutch prefixes and suffixes, the prefixes *her-* and *ver-* are shown as suffixes. Page 208 cites "Baayen (1995)," which is not listed in the bibliography (the reference intended may be to the item listed as 1994b). In Table 6.1 (page 211) and the associated Figure 6.9 (page 212), there are mistakes in the codes for different literary works. (In the table, Emily Brontë's *Wuthering Heights* is coded identically to L. F. Baum's *Tip Manufactures a Pumpkinhead*—surely an implausible confusion—but *Wuthering Heights* seems to be "B1" in the figure; two novels by Arthur Conan Doyle are assigned the same code and identical word lengths in the table, whereas *The Hound of the Baskervilles* is probably the item coded "C2" in the figure.)

The volume is accompanied by a CD-ROM containing numerous relevant software programs; these and various data sets are detailed in a series of four appendices to the book.

**Acknowledgment**

I am grateful to fellow members of the Sussex University probabilistic parsing reading group for insights gained during the weeks when the book under review was our target volume. Responsibility for errors and inadequacies in this review is entirely mine.

## References

Carroll, J. B. 1967. On sampling from a lognormal model of word frequency distribution. In Henry Kučera and W. Nelson Francis, editors, *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI, pages 406–424.

Honoré, A. 1979. Some simple measures of richness of vocabulary. *Association of Literary and Linguistic Computing Bulletin*, 7(2):172–179.

Khmaladze, E. V. 1987. The statistical analysis of large numbers of rare events. Technical Report MS-R8804, Department of Mathematical Sciences, Centrum voor Wiskunde en Informatica. Amsterdam: Centre for Mathematics and Computer Science.

Naranan, S. and V. Balasubrahmanyan. 1998. Models for power law relations in linguistics and information science. *Journal of Quantitative Linguistics*, 5:35–61.

Orlov, J. K. and R. Y. Chitashvili. 1983a. Generalized Z-distribution generating the well-known "rank-distributions." *Bulletin of the Academy of Sciences, Georgia*, 110:269–272.

Orlov, J. K. and R. Y. Chitashvili. 1983b. On the statistical interpretation of Zipf's law. *Bulletin of the Academy of Sciences, Georgia*, 109:505–508.

Sichel, H. S. 1975. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70:542–547.

*Geoffrey Sampson* is Professor of natural language computing at the University of Sussex. Much of his research has concerned statistical parsing techniques; he has contributed the articles on "Statistical Linguistics" to successive editions of the *Oxford International Encyclopedia of Linguistics*. Sampson's address is School of Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton BN1 9QH, England; e-mail: geoffs@cogs.susx.ac.uk.

## Data-Driven Techniques in Speech Synthesis

**R. I. Damper (editor)**
(University of Southampton)

Boston: Kluwer Academic Publishers,
2001, xviii+316 pp; hardbound, ISBN
0-412-81750-0, $145.00, €148.00, £100.00

*Reviewed by*
*Thierry Dutoit*
*Faculté Polytechnique de Mons*

Never say “never.” In 1997, most experts would have sworn that text-to-speech (TTS) synthesis technologies had reached a plateau, from which it would be very hard to leave. Five years later, speech synthesis has been widely and unexpectedly revolutionized by data-driven techniques. Wherever handcrafted rule-based systems were chosen for their incremental design and analytic controllability, machine learning (ML) techniques are now increasingly used for their scalability and genericity, key elements for the design of multilingual, and possibly embedded, TTS systems. The established, “linguist-friendly” paradigm (“if you don’t get a substantial quality increase with ML, stick to expert systems”) is thus being turned into a more pragmatic strategy (“even if it brings a small percentage of error increase, go for ML”). This 316-page book, edited by Robert I. Damper and written by top specialists in the field, addresses such recent advances in data-driven techniques for speech synthesis, with a very strong emphasis on the use of ML techniques for natural language processing issues (and even more specifically for automatic phonetization).

After Damper’s introduction to the architecture of TTS systems in chapter 1, Ghulum Bakiri and Thomas G. Dietterich open a series of seven chapters devoted to automatic grapheme-to-phoneme (GTP) transcription. Their chapter 2, “Constructing High-Accuracy Letter-to-Phoneme Rules with Machine Learning,” examines extensions to NetTalk and NetSpeak, the pioneering (but rather deceiving) work of Sejnowski and Rosenberg. They point out how, by modifying the original multilayer perceptron, it is possible to reach better transcription rates than those possible using established rule-based systems. In chapter 3, “Analogy, the Corpus and Pronunciation,” Kirk P. H. Sullivan presents the idea of pronunciation-by-analogy and its relation to a psychological model of oral reading. The chapter ends with a (somewhat confused) discussion of an implementation of the Sullivan and Damper method for English, Maori, and German. Helen Meng examines the use of probabilistic formal grammars for phonetizing words in chapter 4, “A Hierarchical Lexical Representation for Pronunciation Generation.” Based on a multilevel linguistic description of words that is obtained with a handcrafted context-free grammar, the method attaches probabilities to sibling-sibling transitions in the rules of the parser. Chapter 5, “English Letter–Phoneme Conversion by Stochastic Transducers,” by Robert W. P. Luk and Robert I. Damper, is devoted to the use of stochastic finite-state transducers for GTP conversion in English, a hot but complex topic. After a discussion on maximum-likelihood transduction and on possible ways of achieving automatic GTP alignment (a prerequisite for most GTP transcription systems), it is shown that the best results are obtained when a priori linguistic information is used for alignment. This chapter is dense and thus not truly self-contained.

Sabine Deligne, François Yvon, and Frédéric Bimbot focus on their multigram approach in chapter 6, “Selection of Multiphone Synthesis Units and Grapheme-to-

Phoneme Transcription Using Variable-Length Modeling of Strings," for estimating the probability of a string seen as the concatenation of (automatically derived) independent variable-length sequences of symbols. After presenting the classical multigram approach and its extension to joint multigrams (i.e., on several nonsynchronized streams of symbols), the authors propose two applications for TTS synthesis: that of deriving the set of most frequently needed multiphone units for the design of a concatenative speech synthesis system (which obviously deserves further investigation) and that of performing joint multigram-based GTP conversion. Lazy, or memory-based learning is the subject of chapter 7, "TREETALK: Memory-Based Word Phonemisation," by Walter Daelemans and Antal Van den Bosch. The authors present "normal" lazy learning (IB1-IG), their information-theoretic IGTree-building technique, and a hybrid TRIBL method for optimizing transcription speed while maintaining low error rates. The chapter ends with an analytic discussion on the use of monolithic versus modular GTP systems and surprisingly shows that the best results are obtained when the intermediate levels are left implicit. Chapter 8, "Learnable Phonetic Representations in a Connectionist TTS system—Text to Phonetics," by Andrew D. Cohen, concludes this GTP-oriented part of the book, with a journey into the land of nonsegmental phonology. Departing from the traditionally phoneme-oriented interface between GTP and speech synthesis, a more phonetic interface is examined, which is moreover obtained in an unsupervised way by training a combination of neural networks on a database composed of words in their written and oral forms. The machine itself proposes phonetic units, in the form of attractor basins in a self-organizing map. This chapter, together with chapter 12 by the same author, is certainly one of the most complex and experimental of the book (together they constitute a dense summary of the author's doctoral dissertation).

The four last chapters explore, although to a much lesser extent, the use of data-driven approaches for prosody generation and speech signal synthesis. Chapter 9, "Using the Tilt Intonation Model," by Alan W. Black, Kurt E. Dusterhoff, and Paul A. Taylor, summarizes the authors' Tilt model of intonation. After presenting the easy F0-to-Tilt and Tilt-to-F0 pathways, it is shown that classification and regression trees (CARTs) can do a good job when asked to decide the value of Tilt parameters, using a linguistic prediction feature set. In Chapter 10, "Estimation of Parameters for the Klatt Synthesizer from a Speech Database," John Coleman and Andrew Slater provide a "Klatt synthesizer primer" in which they show how to synthesize high-quality, formant-based English sounds by using automatic acoustic analysis of real speech combined with "tricks of the trade." In Chapter 11, "Training Accent and Phrasing Assignment on Large Corpora," Julia Hirschberg summarizes the use of CART techniques for predicting accent and phrasing assignment (a prerequisite for intonation and duration generation); the method is based on the Pierrehumbert hierarchical description of intonation. The author gives analytic results on several databases (citation-form sentences, news stories by a single speaker, multispeaker broadcast radio and multispeaker spontaneous speech) and obtains results comparable to those derived from a handcrafted rule-based system. The chapter ends with experiments on using text corpora annotated by native speakers in place of time-consuming speech corpora, which make it possible to train models in a (small) fraction of the time needed in the original speech-based training. The book concludes with a short proposal, chapter 12, for extending the ideas of Cohen's first chapter to concatenative speech signal synthesis itself. Cohen proposes a complex combination of neural networks for producing sequences of linear predictive coding (LPC) coefficients and F0 values from the output of his unsupervised GTP system.

I read this book with great pleasure and undoubtedly learned from it. I have no doubt that postgraduate students and researchers in the area will benefit from its reading. It should be clear, however, that prior exposure to neural networks, statistical language modeling, and finite-state models is required to take full advantage of the book, especially for chapters 5–8 and 12. Although most of the material presented in this book appears elsewhere (the authors of each chapter are also their main protagonists and have thus already published their work in various journal papers), it has been given a compact and comprehensive form here.

The book inevitably suffers from "edited book syndrome." The introductions of the first seven chapters tend to have strong overlaps, and the chapters in general contain only few cross-references. Not all chapters are of equal interest for the same person. Researchers will be more interested in chapters 3, 5, and 6, whereas system designers will probably prefer chapters 7, 9, and 11. On the other hand, chapters can be read in virtually any order (except for chapter 1, which should be read first, and chapter 12, which assumes prior reading of chapter 8).

The reader always wants more: One would certainly have loved to get test data, and example training and testing scripts in an included CD-ROM, especially since the authors discuss their own work. More comparative results (possibly as an "add-on" chapter) would have been welcome too. But as judiciously mentioned by several authors, it is not easy to compare technologies with different training hypotheses and testing procedures.

This raises an additional, and maybe broader, question (in the sense that it addresses the field of data-driven GTP in general): Is speech synthesis (and most particularly GTP conversion) seen as a test bed for ML techniques, or is it considered the problem to solve? When comparing systems, most authors emphasize the pros and cons of the underlying technologies (and comment on their possible extensions to various areas), whereas the title of the book somehow suggests a task-oriented approach. Readers who expect the book to provide keys to designing a full data-driven TTS system will be disappointed by the more scientific and prospective considerations they will find. Those interested in having a clearer picture of ML techniques, tested here on speech synthesis problems, will be rewarded.

One last but important caveat: This book surprisingly contains only partial information on data-driven prosody generation and very little information on what seems to be the hottest topic in the TTS industry these last years: data-driven concatenative speech signal synthesis (sometimes referred to as nonuniform unit (NUU) synthesis). Maybe the title is misleading in that respect: The book is actually strongly biased toward language modeling and even more toward GTP conversion.

Summarizing, this book is clearly a must for post-graduate students and researchers in the area of data-driven phonetization. It is the first to propose in-depth, state-of-the-art information on the topic and to offer a comparative view of the underlying technologies. It therefore brings a fresh perspective to this quickly moving field. It can also be used as a pointer to other aspects of data-driven speech synthesis (namely, prosody and speech signal synthesis), although the reader should be aware that these are only very incompletely covered.

*Thierry Dutoit* has been a professor of circuit theory, signal processing, and speech processing at Faculté Polytechnique de Mons in Belgium since 1993. Between 1996 and 1998, he spent 16 months at AT&T–Bell Labs in New Jersey. He is the initiator of the MBROLA speech synthesis project, the author of a reference book on speech synthesis (in English), and the coauthor of a book on speech processing (in French). He has written or cowritten about 60 papers on speech processing and software engineering. Dutoit is also involved in industrial activities as a consultant. Dutoit's address is Faculté Polytechnique de Mons, MULTITEL-TCTS Lab, Initialis Scientific Park, Avenue Copernic, B-7000 Mons, Belgium; e-mail: thierry.dutoit@fpms.ac.be; URL: tcts.fpms.ac.be/∼dutoit.

## Empirical Linguistics

**Geoffrey Sampson**
(University of Sussex)

London: Continuum (Open linguistics
series, edited by Robin Fawcett), 2001,
viii+226 pp; hardbound, ISBN
0-8264-4883-6, $90.00, £55.00

*Reviewed by*
*Steven Abney*
*AT&T Laboratories–Research and University of Michigan*

Geoffrey Sampson has made significant contributions in the area of corpus linguistics, and this book brings together and updates a number of his essays, mostly from recent years but including work whose original publication dates back as far as 1975. Despite this variety of provenance, the volume has been well edited for consistency both in theme and in style. Sampson's stated aim is to give a coherent presentation of an approach to language that he has enunciated in scattered publications over the years, an approach based on systematically collected corpora of naturally occurring language. This approach is a flavor of corpus linguistics, though Sampson prefers the broader term *empirical linguistics*: He considers the corpus to be the primary and essential tool for the empirical study of language.

The ambitious scope of the term *empirical linguistics* is not an accident. Introductory linguistics textbooks usually present the most fundamental distinction among schools of linguistics as that between the empiricists and rationalists. The history of twentieth-century linguistics, as usually presented, is the story of the paradigm shift from empiricism to rationalism marked by the publication of Chomsky's *Syntactic Structures* (1957). Sampson offers his empirical linguistics as an antithesis to Chomsky's generative linguistics. Indeed, in Sampson's view, a second paradigm shift has already occurred. Though "intuition-based linguistic theorizing has lingered on, in some university linguistics departments," as he puts it, empirical linguistics "began to reassert itself in the 1980s, and since about 1990 has moved into the ascendant."

One should not, however, expect from this book a sweeping, definitive exposition of the empirical linguistic paradigm. In particular, anyone seeking an introduction to recent advances in empirical computational linguistics will be disappointed. Nor is it the popularizing work that will convert the world of generative linguistics to corpus methods. It does not speak to generative linguists in their own terms, and it focuses very much on early generative linguistics: center embedding, Yngve's complexity measures (1960, 1961), Katz and Fodor's semantic marker theory (1963), Chomsky's logical structures of linguistic theory (1955 [1975]).

Rather, this book represents a particular concrete example of corpus linguistic investigation, accompanied by a critique of generative linguistics. As such, it provides some fascinating data and provocative philosophical argumentation. It consists of 10 chapters, not including the introduction. Four are empirical studies, one is mathematical, and five are philosophical.

The empirical chapters focus on depth of embedding. Chapter 2 challenges the long-standing constraint against multiple center embedding. It summarizes several variants of the constraint and presents examples from published texts that violate each variant. Unfortunately, after discarding all previous formulations, Sampson does not offer a more adequate formulation, taking instead an agnostic stance on the very existence of the constraint. Chapter 4 examines a related issue, namely, a hard limit on tree depth proposed by Yngve, and concludes that the lack of deep left recursion is not due to an Yngvean constraint on paths but rather is a consequence of the low probability of choosing left-branching expansions. Chapters 3 and 5 examine the influence of genre and social factors on depth of embedding, taking depth of embedding as a proxy for grammatical complexity. Chapter 3 looks at the effect of genre, agreeing with the common wisdom that there is a significant difference in sentence length between technical prose and fiction but concluding that it is not a consequence of a difference in overall structural complexity but is almost entirely ascribable to a difference in the number of immediate constituents in the noun phrase. Chapter 5 examines the hypothesis that social class, age, and gender have an effect on grammatical complexity finding a significant correlation only with age, not with gender or social class. Moreover, Sampson argues that there is a lifelong pattern of increasing grammatical complexity, and takes this as evidence against the existence of a "critical period" of language acquisition ending at puberty.

The mathematical chapter (chapter 7) is rather an outlier in tone and contents. Originally coauthored with William Gale, a statistician at AT&T Bell Laboratories, it gives an excellent exposition of the Good-Turing smoothing method. It presents a step-by-step recipe for computing Good-Turing discounts and an accessible but by no means trivializing account of the theory behind the method.

The remaining five chapters are more concerned with the philosophy of linguistics than with empirical investigation. Chapter 6 advocates a greater emphasis on taxonomy (particularly in the form of treebanks) in linguistics. Appeal is made to the example of biology, in which the systematizing work of Linnaeus was an essential preliminary for modern biological theory. Chapter 8 scoffs at the use of linguistic intuitions and invented examples, as opposed to corpus data, comparing a linguist using intuitions to a meteorologist who theorizes on the basis of intuitions about weather forecasting. Chapter 9 is something of an interlude, being a specific and quite detailed attack on Chomsky's *Logical Structure of Linguistic Theory*. It is perhaps best summarized by quoting a passage: "Alan Sokal and Jean Bricmont [1998] have recently documented the way that a number of twentieth-century 'intellectuals' created an air of profundity in their writings on social and psychological topics by using mathematical terminology and symbols which, to readers ignorant of mathematics, look impressive, though to professional mathematicians they are nonsensical." Chapter 10 claims that the notion of ungrammatical sentence is a Chomskyan invention, the traditional statements on the matter being in the mode "this sentence cannot be used that way," not "this sentence cannot be used." Sampson also challenges the notion of a fixed grammar, arguing that there are no invalid expansions for any category, merely a long tail of low-frequency expansions. Finally, in chapter 11, he argues that some aspects of language are beyond the limits of science. Rejecting first Katz and Fodor's formalism for lexical semantics, he goes on to reject the idea "that words have definite meanings capable of being captured in symbols of *any* formal notation" and argues that learning word meanings is more like learning to dress fashionably: There is no truth to the matter, one just tries to imitate what the more authoritatively fashionable do.

In sum, the book will appeal most to those who are interested in constraints on depth of embedding, to those interested in corpus linguistics, and to those interested in criticism of generative linguistics, particularly early generative linguistics. Anyone implementing Good-Turing smoothing will also find chapter 7 useful. It is to be recommended not as a general introduction to modern empirical linguistics, but as an exposition and example of a particularly pure strain of linguistic empiricism. To my mind, it also reveals the weaknesses of pure empiricism, especially the lack of elucidation of mechanisms giving rise to phenomena, but it certainly cannot be accused of compromising its principles.

## References

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton, The Hague.

Chomsky, Noam. [1955] 1975. *The Logical Structure of Linguistic Theory*. Plenum, New York.

Katz, Jerrold J. and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language*, 39:170–210.

Sokal, Alan and Jean Bricmont. 1998. *Intellectual Impostures: Postmodern Philosophers' Abuse of Science*. Profile. (Published in the U.S. under the title *Fashionable Nonsense*.)

Yngve, Victor H. 1960. A model and a hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466.

Yngve, Victor H. 1961. The depth hypothesis. In Roman Jakobson, editor, *Structure of Language and Its Mathematical Aspects*, volume 12 of *Proceedings of Symposia in Applied Mathematics*. American Mathematical Society, Providence, RI, pages 130–138.

*Steven Abney* wrote this review while a principal member of the research staff at AT&T Laboratories–Research; he is now associate professor of linguistics at the University of Michigan. His interests include parsing, language learning, stochastic models, corpus methods, syntax, and semantics. Abney's address is Department of Linguistics, University of Michigan, Ann Arbor, MI 48109; e-mail: spa@vinartus.net; URL: www.vinartus.net/spa.

# Computational Nonlinear Morphology with Emphasis on Semitic Languages

**George Anton Kiraz**
(Beth Mardutho: The Syriac Institute)

Cambridge: Cambridge University
Press (Studies in natural language
processing, edited by Branimir
Boguraev and Steven Bird), 2001,
xxi+171 pp; hardbound, ISBN
0-521-63196-3, $60.00

*Reviewed by*
*Markus Walther*
*Panasonic Speech Technology Laboratory*

## 1. Introduction

Computational morphology would be an almost trivial exercise if every language were like English. Here, chopp-ing off the occasion-al affix-es, of which there are not too many, is sufficient to isolate the stem, perhaps modulo a few (morpho)graphemic rules to handle phenomena like the consonant doubling we just saw in *chopping*. This relative ease with which one can identify the core meaning component of a word explains the success of rather simple stemming algorithms for English or the way in which most part-of-speech (POS) taggers get away with just examining bounded initial and final substrings of unknown words for guessing their parts of speech. In contrast, this book outlines a computational approach to morphology that explicitly includes languages from the Semitic family, in particular Arabic and Syriac, where the linearity hypothesis—every word can be built via string concatenation of its component morphemes—seems to break down (we will take up the validity of that assumption below).

Example 1 illustrates the problem at hand with Syriac verb forms of the root {q$_1$ṭ$_2$l$_3$} 'notion of killing' (from Kiraz [1996]).

| (1) | | *Stem shape* | *Form* | *Morphs* | *Gloss* |
|---|---|---|---|---|---|
| | *a.* | $C_1C_2V_2C_3$ | *qṭal* | $a_1a_2$ *past act.* | *he killed* |
| | *b.* | | *neqṭol* | *ne- 3 sg. m.,* $a_1o_2$ *fut.* | *he will kill* |
| | *c.* | | *ʔ̣ eθqṭel* | *ʔ̣ eθ- refl.,* $a_1e_2$ *past pass.* | *he was killed* |
| | *d.* | $C_1V_1C_2C_3$ | *qaṭleh* | $a_1a_2$ *past act., -eh obj.* | *he killed-OBJ* |
| | *e.* | $C_1C_2C_3$ | *neqṭluː n* | *ne- -u ː n 3 pl. m.,* $a_1o_2$ *fut.* | *they (m.) will kill* |

Notice the use of subscripts as a visual aid in pairing up abstract consonantal (C) and vocalic (V) stem positions with concrete segments. The stem shapes show how root and tense/aspect morphemes are interdigitated. Also evident is the considerable variability in stem vowel (non)realization, leading to vowelless stems in the extreme case (1e).

## 2. Content

The opening chapter begins by specifying the intended wide audience, namely computational, theoretical, and applied linguists as well as Semitists. It then addresses linguistic preliminaries, including brief introductions to morphology and autosegmental phonology, before proceeding to some formal language theory and unification. Introductory applications of these to selected morphology and phonology problems are given. Very briefly, the bare basics of Semitic noun and verb morphology are touched upon as well as some peculiarities of its predominant writing system.

Chapter 2 is a very useful survey of three mainstream approaches to the formal description of Semitic word formation that differ in terms of which units form the template, that is, the central sequencing device (CV vs. moraic), and how many templates are assumed (affixational approach). Here Kiraz strictly focuses on pre-optimality-theoretic work by John McCarthy and Alan Prince, two influential theorists in generative linguistics. Notably the author also draws attention to aspects of Semitic morphology beyond the stem, highlighting the existence of various affixation processes as well as phonological effects such as vowel deletions sensitive to syllable structure.

Chapter 3 begins by mentioning the work of Kaplan and Kay (1994) on cascaded finite-state rules but mostly focuses on further developments of the two-level model (Koskenniemi 1983) for parallel rule application in a finite-state setting, since Kiraz intends to use an extended formalism from that class. Among the modifications reviewed are mapping of sequences rather than single symbols only, unequal-length mappings, unification over finite-valued features, and proper treatment of obligatory rules.

Chapter 4 prepares the ground for Kiraz's own work by reviewing no less than nine different approaches to Semitic computational morphology. They broadly fall into two classes, one following the autosegmental, multitiered approach, whether expressed by mappings between several automaton tapes or intensional descriptions that codescribe a single tape. The other class follows no particular theory but often uses regular set intersection to combine root, template, and vowel pattern.

The central Chapter 5 finally introduces Kiraz's own multitier formalism. Here we find comprehensive descriptions and formal definitions of the lexicon and rewrite-rule components. The former consists of sublexica corresponding to the various lexical tiers or tapes, whereas the latter allows two-level-style context restriction and surface coercion rules. All the modifications discussed in chapter 3 are incorporated here, and proposals for handling morphotactics are described as well.

Chapter 6 now applies the multitier formalism to selected problems of Arabic morphology. It details the three approaches of chapter 2 to verb stem formation, giving formal rules and lexicon entries that allow the reader to simulate sample stem derivations in Kiraz's framework. With regard to noun morphology, "broken" plurals like *xaatam* 'signet-ring (sg.)' ∼ *xawaatim* '(pl.)' receive a formal analysis as well. Kiraz discusses issues of nonlinearity versus linearity and generation of partially voweled spellings before finishing the chapter with a rule-based treatment of glyph alternations in Syriac script.

Chapter 7 develops the compilation of Kiraz's formalism into multitape automata, broadly using the concepts and methodology of Kaplan and Kay while introducing additional regular operators for *n*-way regular relations. Because the different stages can get quite involved technically, they are illustrated step by step with the help of simple examples and automaton diagrams.

The book concludes in Chapter 8 by first presenting a short discussion of applications of the formalism to general autosegmental problems, illustrated with an

African tone language. Then it touches on the subjects of disambiguation of Semitic orthographic representations (high ambiguity due to absence of short vowels), semantics in Semitic (sense disambiguation), and productivity (mainly extension of existing roots to previously unused patterns). Interestingly, Kiraz speculates that addressing productivity might involve weighted automata to express the preference for roots to attach to lexically known patterns without completely ruling out a new-word interpretation.

Finally, five pages of references and three indices are provided. The book appears to be carefully edited, has a professional layout, and is remarkably free of typographic and spelling errors.

## 3. Critique

The author stresses (p. xv) that the research for this book, originally his Ph.D. thesis, took place between 1992 and 1996. With five years to publication, there is considerable risk of new developments in the field (or a revival of old ideas) that could provide competing insight or weaken central claims. This section will discuss some of the more problematic aspects of this book in this regard.

But first, what about its suitability for the stated target audience? Although bridging the gap between the separate disciplines that share an interest in the subject is certainly a laudable goal, this reviewer is quite unsure whether the book succeeds in meeting it. The Semitist will probably feel overwhelmed by the amount of mathematical formalism, without getting rewarded in the end by, say, application to interesting comparative or diachronic problems from his field of interest. Theoretical linguists will in addition recognize immediately that the book does not cover constraint-based approaches like optimality theory, which from the very beginning were strongly motivated by prosodic morphology (McCarthy and Prince 1993), of which the Semitic kind is a fine example. If merely adapting now-abandoned analyses to a computational setting is not a particularly strong selling point for this group, then neither is the absence of a detailed treatment of some of the more interesting issues that Semitic presents, such as how to capture its morphological richness with few parameterized or prioritized principles, how to regularize the apparent irregularity of weak verbs, and so on. That leaves the computational linguist who wants to, say, build a practical morphological analyzer for Arabic or understand the minimal computational requirements for a plausible model of morphology that includes Semitic languages.

Following the recent trend toward data-intensive, empirically oriented computational linguistics, such a reader will probably first want to see a decent introduction to the phenomena at hand. But what they get is rather disappointing. Kiraz does describe the Arabic "broken" plural, giving a number of example pairs, but without proper discussion of its productivity and the corresponding "sound" plural it is a bit hard to understand why it is worth being modeled by rules instead of lexical listing. For verbs, no exemplary paradigms of surface forms are given at all, and no tables list nontrivial excerpts of the morphological system of a language as unfamiliar as Syriac. When Arabic stems are presented (page 34), the reader has to wait 28 pages to be informed that, actually, the form */nkutib/* is pronounced [*ʔinkutib*] 'write (measure 7, pass.)'. Of course, this makes a huge difference: The former is prosodically ill formed, unlike the latter, whose prefix *ʔin-* is a well-formed syllable. Insightful linguistic analysis is hardly possible when using defective data, yet Kiraz bases his formal analysis on them (page 104f). Regrettably we are often not given enough detail about the prosodic systems of both languages: Avoidance of initial CC clusters in Arabic is mentioned in passing, but is it exceptionless? And what about the distribution of the same in Syriac, where such clusters are allowed? In a section on neologisms (page 152), only the expert will not be puzzled when Kiraz cites two such forms without glosses; one cannot even pronounce the Syriac form of the two because the transliterated vowel symbol å is not explained (page xx). In sum, the nonspecialist is given too little of the big picture to be able to come up with alternative ideas about plausible models for the data.

Next, the reader may start wondering whether it is actually true that "[u]sing the nonlinear model is *crucial* for developing Semitic systems" (page 110, emphasis added). Kiraz himself never questions the tradition that interprets the conceptual autonomy of consonantal root, template, and vowel sequence[1] as technical nonlinearity.

He does show, however, that actually a nonlinear representation is harmful everywhere but in the stem, for example, leading to duplication of rules when coverage is extended to affixed forms (page 112f). As a consequence, he must weaken his architecture to provide a second stage in which rules postprocess fully linearized verb stems; the same setup is proposed for broken plural formation in nouns, because vowel length and prosodic shape transfer from singular to plural and cannot be read off the components alone. A third, again linear, stage optionally deletes short vowels from the fully pronounceable surface form to map to partially voweled orthographic representations. At this point good scientific reductionism would seem to suggest trying to reduce nonlinearity to zero, but Kiraz offers no discussion of why any such alternative won't fly.

In fact, such an alternative has been proposed by Hudson (1986) for Arabic. In the briefest of sketches, a modernized version taken from Walther (1999) goes like this: We replace object strings by partial descriptions and encode stems with the help of optionality parentheses for zero-alternating vowels, for example, *q*(*a*)*t*(*o*)*l* for the future stem. While one such description denotes four surface strings, nonalternating affixes are represented without optional segments, giving *neq*(*a*)*t*(*o*)*l* after concatenation (cf. 1b). Using the central insight that the shape of entire word forms, not stems alone, is governed by syllable structure constraints, here (C)CV(V)(C), we are left with the set {*neqatol*,*neqtol*}. Assigning a weight to every realized vowel, we can finally apply a left-to-right greedy shortest-path algorithm to correctly prefer *neqtol* over *neqatol* because it omits an alternating stem vowel as early as possible. Note that left-to-right incrementality is psycholinguistically plausible and leads on average to an earlier recognition point for the root. This approach, which has been used to formulate sizeable morphological grammars for Tigrinya (Walther 1999) and Modern Hebrew (Walther 1998), can also be implemented in finite-state terms. With the aid of an inheritance-based formalism, redundancy in stem descriptions would be kept minimal, thus retaining a logical, but not object-level, autonomy of stem components while accommodating exceptions at the same time. In contrast to Kiraz's approach, which must employ baroque vowel deletion rules that operate right to left to edit the abstract stems under affixation, the constraint-based alternative sketched is much more explanatory in terms of why Semitic stems exhibit so much shape *variance* instead of the shape invariance predicted by Kiraz's rigid templates: They simply respond to both the language-particular restrictive syllable canon and universal demands for processing economy. Under this perspective, Semitic morphology is formally atemplatic

[1] Although these are usually motivated both by descriptive economy and identifiable semantic contribution, Kiraz does not discuss the significant extent to which stems in Semitic languages like Modern Hebrew have noncompositional meanings that cannot be predicted from their components.

and concatenative, differing mainly by its regular use of vowel/zero alternation and ablaut (cf. sporadic cases like German *Segel* ∼ *Segl-er* 'sail ∼ sailor' and English *sing* ∼ *sang* ∼ *sung*).[2]

If its linguistic motivation is found wanting, perhaps the main strength of Kiraz's proposal comes from the technical side, with greater computational efficiency and just the right expressivity? In fact, this is what Kiraz seems to have in mind (pages 68, 111). When discussing related work that dispenses with multiple lexical tapes or tiers—while still sharing the template idea—he identifies intersection-based and mapping-based approaches as the main players. Simply put, in the former, consonantally underspecified template automata like *CaCaC* are intersected with vocalically underspecified root automata such as *kVtVb*, whereas in the latter, one rewrite rule is constructed per stem that specifies the linear arrangement of its components at compile time. In his critique Kiraz alleges that intersection loses bidirectionality; that is, parsing cannot reliably recover the root and the other components if given just stems, that one-rule-per-stem is highly redundant, and that both approaches are computationally expensive at compile time compared to his multitape approach.

Just as Kiraz modifies traditional automata, however, so can proponents of the intersection approach (which is similar to the alternative outlined above). For example, to recover whether a segment originates from root or vocalic pattern or affix, one could envision composite labels ⟨*segment*, *origin*⟩ on automata transitions, where *segment* parts match traditionally, whereas *origin*s are unioned together. The parse string would start out with empty *origin* sets.

As for the other advantage, compile-time efficiency: This is a notoriously risky argument, given that computers get faster all the time and that the main attractiveness of finite-state processing lies in its fast *run-time* behavior. In this regard it is curious that Kiraz cites his paper (Kiraz 2000) but does not incorporate the empirical evaluation from that paper into the book to strengthen his claim. In any case, recent results by Beesley and Karttunen (2000) show that fast compilation no longer necessitates a multitier model. Their proposal is that automata strings could themselves contain textual representations of regular expressions, with a new *compile-replace* operator allowing for in situ evaluation and substitution. This reduces compile time from hours to a few minutes for a large-scale Arabic morphology, using compile-replace for stem formation and composition with finite-state rule transducers to map to the surface forms. Although this might seem like an eclectic mix of different strategies, recall that Kiraz himself has a hybrid system with several composed stages: Does this imply that his multitier formalism by itself is not expressive enough for practical grammars?

To be sure, the book does have its strong sides, including good reviews of related work and an exposition of a particular multitape finite-state formalism that is detailed enough to allow the interested reader to implement it, and—if so desired—create a working morphology system for Semitic and other languages. Therefore I would recommend it as a useful source of inspiration for researchers in the field, as long as their foreseen applications are unaffected by the criticism presented above.

[2] Kiraz (1996) defends the necessity of abstract stems such as Syriac **katab* because a rule that turns certain plosives into same-place fricatives applies after short vowels (→ **kaθav*), which may be deleted in surface forms (→ *kθav* 'he wrote'). In fact, however, a surface-true prosodic reformulation *can* account for his data: Those plosives fricativize after noncoda segments, here the complex-onset member *k* and nucleus *a*.

## References

Beesley, Kenneth R. and Lauri Karttunen. 2000. Finite-state non-concatenative morphotactics. In Jason Eisner, Lauri Karttunen, and Alain Thériault, editors, *Proceedings of the Fifth Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON-2000)*, pages 1–12, Luxembourg.

Hudson, Grover. 1986. Arabic root and pattern morphology without tiers. *Journal of Linguistics*, 22:85–122.

Kaplan, Ronald and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378.

Kiraz, George Anton. 1996. Syriac morphology: From a linguistic model to a computational implementation. In R. Lavenant, editor, *VII Symposium Syriacum 1996*. Orientalia Christiana Analecta, Rome.

Kiraz, George Anton. 2000. Multitiered nonlinear morphology using multi-tape finite automata: A case study on Syriac and Arabic. *Computational Linguistics*, 26(1):77–105.

Koskenniemi, Kimmo. 1983. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, University of Helsinki, Helsinki.

McCarthy, John and Alan Prince. 1993. Prosodic morphology I: Constraint interaction and satisfaction. Technical Report RuCCS-TR-3, Rutgers University Center for Cognitive Science.

Walther, Markus. 1998. Computing declarative prosodic morphology. In Mark Ellison, editor, *Proceedings of the Fourth Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON 98)*, pages 11–20, Montreal.

Walther, Markus. 1999. *Deklarative prosodische Morphologie: Constraint-basierte Analysen und Computermodelle zum Finnischen und Tigrinya*. Niemeyer, Tübingen, Germany.

*Markus Walther* has published on computational phonology and morphology using logic-based and finite-state formalisms, with applications including reduplication and Semitic word formation in Tigrinya and Modern Hebrew. He now works in research and development for text-to-speech synthesis. Walther's address is Panasonic Speech Technology Laboratory, 3888 State Street #202, Santa Barbara, CA 93101; e-mail: mwalther@stl.research.panasonic.com; URL: www.markus-walther.de.