Briefly Noted

English for the Computer: The SUSANNE Corpus and Analytic Scheme

Geoffrey Sampson
(University of Sussex)

Oxford: Clarendon Press, 1995,[1] ix+499 pp;
hardbound, ISBN 0-19-824023-6, $115.00,
£79.00

Over the past 10–20 years, there has been increasing interest in grammatical / syntactic annotation schemes for corpora. Annotated corpora are essential for training and testing taggers and parsers, for describing the use of lexical and grammatical features, and for comprehensive analyses of registers or sublanguages. Several annotation schemes have been developed over this period, including both tagsets (see the survey by Leech [1997]) and parsing schemes for syntactic treebanks (see the survey by Leech and Eyes [1997]).

The SUSANNE analytic scheme is probably the most detailed and explicit of these annotation frameworks. In this 500-page book, Sampson describes every detail of the scheme, which is based on a multiyear analysis and annotation of the SUSANNE Corpus.

This book has different goals from most corpus-based studies. It does not present the results of corpus analysis or describe new approaches or tools for natural language processing. Instead, the goals here address the need for a standard. At present, every research group has its own set of standards for corpus annotation, resulting in unnecessary preprocessing when corpora are exchanged across sites, and making it difficult to compare evaluations of corpus analysis tools. (For example, it is notoriously difficult to compare accuracy rates for taggers because they vary so widely in the range of phenomena annotated.) Sampson proposes a comprehensive annotation scheme, covering part-of-speech tagging, surface grammar, and issues of underlying "logical grammar," with the hope that the scheme might provide the basis for a widely adopted standard.

To develop the annotation scheme, Sampson and his colleagues manually annotated the 130,000-word SUSANNE Corpus. This corpus has had several lives. It began as a subset of the Brown Corpus, selected from four text categories: press reportage, belles lettres, "learned" prose, and adventure fiction. That subcorpus was manually analyzed by Alvar Ellegård and colleagues at Gothenburg University and came to be known as the Gothenburg Corpus. Sampson's team then analyzed this same subcorpus in greater detail, together with a small sample of speech from the London-Lund Corpus, resulting in the SUSANNE Corpus.

The book under review documents the hundreds of decisions required for this annotation process. This resource should prove useful for researchers faced with these same decisions during corpus analysis. However, the level of detail will sometimes overwhelm computational and corpus linguists attempting to develop automatic taggers and parsers. Many of these decisions clearly require a human analyst, and thus it is not clear how useful they will be for machine processing. Others end up seeming arbitrary, despite the detailed justifications. For example, *Flagstaff*—my home town—is tagged as a common noun (page 87), because the word originally referred to a flagpole. In contrast, *Greenwood* is tagged as a proper noun (page 88), because it was originally the surname of a colonist. These principles seem clear, even if nearly impossible to implement automatically. However, I had trouble reconciling those decisions with the tagging of *Red* as a proper noun (masculine forename) in *Red Hogan* but as an adjective in *V. E. (Red) Berry* (page 138).

I suspect that Sampson would not be bothered by such criticisms. He did not set out to develop the perfect annotation scheme, and he readily agrees that his scheme—like all others—has controversial points. What makes this scheme unique is its attempt to be comprehensive and explicit about all details. In this respect, corpus researchers will find themselves relying on this reference book increasingly as they work toward the goal of readily exchangeable resources.—*Douglas Biber, Northern Arizona University*

[1] Received for review, May 2001.

References

Leech, Geoffrey. 1997. Grammatical tagging. In Roger Garside, Geoffrey Leech, and Tony McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pages 19–33.

Leech, Geoffrey and Elizabeth Eyes. 1997. Syntactic annotation: Treebanks. In Roger Garside, Geoffrey Leech, and Tony McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pages 34–52.

Mathematical Foundations of Information Retrieval

Sándor Dominich
(University of Veszprém)

Dordrecht: Kluwer Academic Publishers
(Mathematical modelling: Theory and
applications, edited by R. Lowen,
volume 12), 2001, xxi+284 pp; hardbound,
ISBN 0-7923-6861-4, $107.00, £67.00, €99.00

While libraries have long stored informative material for later use, information retrieval as we now know it did not begin to coalesce as a discipline until the 1960s and early 1970s, with advances in commercial systems and influential research publications by scholars including Lancaster, Maron, Salton, and Sparck Jones. While commercial systems of the time most commonly accepted Boolean queries as input, describing the relationships desired between human-assigned index terms in the documents to be retrieved, researchers began developing models and software consistent with term-weighting systems; these evolved into the methods used by today's search engines in ranking documents. While many of the term-weighting and automatic indexing schemes were initially rather simple, they have grown in complexity, based on developments in retrieval and linguistic theory and years of experimentation. While more sophisticated linguistic methods have been studied in retrieval contexts nearly as long as retrieval itself has been studied, the relative level of satisfaction with the performance of retrieval systems using simple automated indexing has kept the linguistic focus of retrieval researchers on individual terms, usually assuming a "bag of terms" model.

Dominich summarizes many of the mathematical foundations of various information retrieval models. Chapter 2 provides the core mathematical material in the book. A wide range of concepts is presented, with a section for each of the following: logic, set theory, relations, functions, families of sets, algebra, calculus, differential equations, vectors, probability, fuzzy sets, metric spaces, topology, graph theory, matroid theory, recursion and complexity theory, and artificial neural networks. The sections are typically broken down into formal definitions, theorems, and examples. The definitions and theorems are clear and relatively easy to understand. Those seeking longer or deeper mathematical expositions on these topics will need to go to the mathematical literature; however, in most cases, the material provided by Dominich will be adequate for linguists and retrieval specialists trying to understand retrieval models. The chapter has little on the relative strengths, weaknesses, and consequences of the adoption of particular mathematical paradigms, which may be frustrating to those asking "Why?" The numeric or symbolic examples provided at the ends of many sections are brief but very useful.

Chapter 3 addresses retrieval models, with one-third of the chapter addressing traditional text retrieval models (Boolean, vector, and probabilistic), one-third addressing "nonclassical" models that have yet to see much commercial use but would be of interest to philosophers and linguists (e.g., ideas based on the works of Barwise and Devlin), and one-third addressing "alternative" models of information retrieval, including a presentation on latent semantic indexing (four pages) and natural language processing (one page). Other brief discussions of techniques using linguistic information are spread throughout the chapters.

Chapter 4 addresses how information retrieval, as modeled in Chapter 3, may be based rigorously on the mathematics presented in Chapter 2. Much of this constitutes original research. It is helpful to see the work of a variety of authors brought together into the uniform presentation provided by Dominich. Chapter 5 formally examines retrieval effectiveness. This presentation goes into far more depth than do most information retrieval books and provides a helpful summary of the foundations of the area. The appendices contain algorithms and MathCAD code for several retrieval models.

The book uses a section-numbering style where the chapter number isn't at the beginning of each section number, nor is the chapter number in the header or the footer for most pages. This makes navigating through the book unnecessarily difficult.

Mathematical Foundations of Information Retrieval will be useful to those computational linguists who want an accessible yet mathematically rigorous presentation of the foundations of information retrieval algorithms and models.—*Robert Losee, University of North Carolina*