COMPUTATIONAL APPROACHES TO THE LEXICON

Edited by

B. T. S. ATKINS AND
A. ZAMPOLLI

CLARENDON PRESS OXFORD

2 Computational Approaches to the Lexicon: An Overview

B. T. S. ATKINS, BETH LEVIN,
and A. ZAMPOLLI

Computational lexicology and lexicography, as the name implies, is a field with close links to other disciplines, in particular computer science, theoretical and computational linguistics, and theoretical and practical lexicography. It is a subject which over the past twenty-five years or so has acquired a personality of its own, with its own growing group of specialists and its own body of literature. Researchers in this new field come from all the contributing disciplines, as reflected, for instance, in the authorship of this volume, which offers a broad perspective on the work being done in the principal areas of activity, ranging from relevant aspects of linguistic theory, through the reusability of lexical resources, to machine-assisted dictionary compiling.

In this overview we trace the development of computational lexicology and lexicography up to the beginning of the present decade, when it finally, we believe, achieved recognition as a scholarly discipline in its own right. First, we look briefly at some relevant work in theoretical linguistics, since over the past two decades this field has been characterized by a growing interest in the lexicon. We note how, in parallel with these developments within theoretical linguistics, the introduction of computational techniques into linguistic research gave rise to the new field of computational linguistics and the related and more specific subject of corpus linguistics. Then we outline the expanding role of the computer in lexicography proper. We next survey the availability of lexical resources and the attempts currently being made to ensure their reusability. We touch upon the cross-fertilization from a variety of specialities that is typical of this new field, where the collaboration of workers from different disciplines

Many people have made valuable contributions to this chapter. We should like to thank in particular Robert Amsler, Jeremy Clear, Charles Fillmore, Patrick Hanks, Susan Hockey, Nancy Ide, Stig Johansson, Ole Norling-Christensen, Nicholas Ostler, Krista Varantola, Edmund Weiner, and Annie Zaenen. We are also grateful to Olivia Chang for help with the bibliography.

(reflected in several of the chapters in this volume) has, we believe, been beneficial to the subject as a whole. Finally, we set the various chapters included in this volume in the context of current work in the field of computational approaches to the lexicon.

1. SOME RELEVANT ASPECTS OF THEORETICAL WORK ON THE LEXICON

In theoretical linguistics work in the late 1950s and 1960s, syntax had come to encompass more and more of the regular aspects of language (Chomsky 1957, 1965), ranging from passivization and question formation (Chomsky 1957, 1965) to word-formation processes such as nominalization (Lees 1960) and compounding (Lees 1960; Roeper and Siegel 1978). The lexicon was considered to be a repository of idiosyncrasies, precisely the view articulated by Bloomfield in 1933, who wrote, 'The lexicon is really an appendix of the grammar, a list of basic irregularities' (Bloomfield 1933: 274). Even certain aspects of word meaning were drawn into the syntax, as in Case Grammar (Fillmore 1968) and generative semantics (McCawley 1973, 1979; Lakoff 1971; Ross 1972).

As early as 1960, however, lexicography attracted the attention of theoretical linguists (Householder and Saporta 1961), and the 1960s and early 1970s produced several works presaging the current interest in the lexicon, among them being the semantic analysis studies of Jerrold Katz and his colleagues (Katz and Fodor 1963; Katz and Postal 1964; Katz 1972), and the lexical listings of Householder and his co-researchers (Householder *et al.* 1964, 1965). During the 1970s, within the generative tradition, the lexicon came to the fore, not only in the study of word formation, but also in the study of a variety of phenomena that had previously been considered to be within the domain of syntax. This shift followed the publication in 1970 of Chomsky's 'Remarks on Nominalizations', which signalled a turning-point in the way the lexicon was viewed within the generative grammar tradition. This paper distinguished between 'syntactic' and 'lexical' phenomena (see also Wasow 1977 for a clear elaboration of this distinction), allowing for the existence of regular processes in the lexicon. In parenthesis, no such clear-cut distinction had been made in non-generative linguistic thinking, such as systemic grammar (Halliday 1966) and other European traditions. Lexical functional grammar (LFG) (Bresnan 1982; Bresnan and Kanerva 1989), for instance, evolved in an effort to deal with many of the processes previously handled by syntactic transformations via lexical rules. In the Chomskyan approach, specifically in the Government-Binding (GB) framework (Chomsky 1981, 1986), as part of the effort to constrain the power of syntactic rules, many syntactic properties of a word came to be viewed as a projection of its lexical properties (Chomsky 1981, 1986; Grimshaw 1981; Pesetsky 1982; Stowell 1981). This view was also shared by LFG and by Generalized Phrase Structure Grammar (GPSG) (Gazdar *et al.* 1985) and its descendant Head-driven Phrase Structure Grammar (HPSG) (Pollard and Sag 1987), as pointed out by Wasow (1985: 202):

> In these three theories . . . clause structure is largely predictable from the semantics of predicates . . . Grammar rules are needed only to state certain language-wide generalizations about how the pieces of sentences are put together and to deal with apparent exceptions to the normal patterns. Most of what was stipulated in the grammars of earlier theories is taken to be a function of lexical semantics.

As a result of the changed view of the lexicon and its relation to syntax, it became important to determine the layout of lexical entries and to explore what needed to be stated in the lexical entry of a word. In this context work began on the MIT Lexicon Project, where the study and recording of lesser-known languages proceeded in tandem with theoretical investigations into lexical representations (Hale and Keyser 1986, 1987; Laughren and Nash 1983; Levin 1993; Rappaport and Levin 1988; Rappaport, Levin and Laughren 1988; among others). The work of the Lexicon Project also built on Jackendoff's important continuing work on the study of the mental representations that serve as a syntax of thought (Jackendoff 1983, 1990), a line of research which Jackendoff began by building on Gruber's earlier attempt to incorporate a lexicon that took into account the semantic properties of lexical items within the transformational grammar model of the 1960s (Gruber 1965).

Recently, then, theoretical linguists in the generative tradition have come to view the lexicon as a legitimate object of study in its own right, closely linked to, but distinct from, the syntax of a language, and the semantic underpinnings of lexical syntactic properties have become a new focus of attention. However, these issues have always been important within certain European traditions. Work in valency grammar has always recognized the primacy of the predicate. This work was further enriched by work in Case Grammar and its descendants: Starosta's (1988) work on lexicase exemplifies this trend, as does the work of Somers (1987).

Studies of linguistic semantics have flourished and continue to flourish outside of work in the generative tradition. Some of these studies have received an additional impetus from a variety of related fields, including anthropology, psychology, and philosophy, as instanced by—among others—the work of Miller and Johnson-Laird (1976). Particularly important in the recently emerging area that has become known as cognitive linguistics has been the work in prototype theory and linguistic categorization of Rosch and her colleagues in psycholinguistics (Rosch 1975; Rosch and Lloyd 1978; Rosch and Mervis 1975), and Berlin and Kay (1969) in anthropological linguistics. More recently, these notions have been incorporated into Lakoff's work on cognitive grammar and semantics, in particular the role of metaphor in word meaning (Lakoff and Johnson 1980), and the concept of radial categories (Brugman and Lakoff 1988; Lakoff 1987). The work of Fillmore, Kay, and other Berkeley linguists has resulted in the continuing development of syntactic and semantic frameworks which allow for mutually compatible descriptions, namely construction grammar and frame semantics (Fillmore *et al*. 1988; Fillmore and Atkins 1992).

Another line of lexical studies has focused on lexical relations—the possible semantic relations between lexical items, including but not limited to the well-studied relations of synonymy, antonymy, and hyponymy—as a way of characterizing their lexical properties. Particularly influential here has been the work of Mel'čuk and his colleagues on the *Explanatory Combinatory Dictionary* (Mel'čuk 1973, 1988; Mel'čuk *et al*. 1981), which was built on an expanded set of lexical relations that goes well beyond those traditionally studied. The notion of lexical relations has also been the basis of Miller's massive and comprehensive WordNet project. This is an attempt to explore the lexical organization of English and the representation of lexical concepts by building an on-line lexical database of English organized around lexical relations, primarily the relation of synonymy (Miller *et al*. 1988; Miller 1990; Miller and Fellbaum 1991).

Thus, at the start of the 1990s we have a situation where linguists from very different traditions are all concerned with the semantic and syntactic properties of words, and the relationships between the two.

2. COMPUTATIONAL LINGUISTICS

The past several decades have also seen considerable activity in the field of computational linguistics (CL). As some of that work was tied to specific linguistic theories (for instance, many parsers have an integral relationship with a particular syntactic framework), the developing interest in the lexicon was carried over into this new field. Small lexicons have often been sufficient in highly theoretical work.

However, with the emergence of the field that came to be known as 'the language industries', natural language processing (NLP) came into its own. Computers were used in applications which range from the most sophisticated machine-translation systems (still far in the future, although useful machine-assisted translation systems already exist, such as SYSTRAN, currently being used in the European Community, or METAL, a more recent system commercialized by Siemens) through speech-to-text and text-to-speech projects, down to handheld pocket electronic dictionaries, and to the spelling checkers used in word-processing programs.

One of the major effects that the growth of the language industries had on computational linguistics as a whole was to introduce the need for large-scale lexicons. Thus NLP systems face what has been called the 'lexical bottleneck' (Byrd 1989*a*, n.d., among others)—limitations in system performance attributable to the need for larger lexicons. It has become of paramount importance to find sources of data and to develop extraction and analysis techniques that allow the building of effective large lexicons with the minimum of effort and expense. These needs sparked a renewed interest in corpus linguistics and in lexicography, as both those fields offered potential help in overcoming the lexical bottleneck.

3. CORPUS LINGUISTICS

Corpus linguistics seeks to further our understanding of language through the analysis of large quantities of naturally occurring data, and the fruits of this work are of immense interest to computational linguists. There is a long tradition of corpus linguistics studies in Europe. Even as early as the end of the nineteenth century large corpora were used to develop the first frequency lists. F. W. Kaeding's study, for instance, entitled *Haufigkeitswörterbuch der deutschen Sprache* (published in 1898) was based on a corpus of eleven million words manually analysed to establish the frequencies of graphemes, syllables, and words. Using this data, J. B. Estoup, in his *Gammes sténographiques* (published in Paris in 1907), noted the statistical regularities in the list of the word forms in a text, ranked in order of decreasing frequency, and this became the starting-point of the well-known work of G. K. Zipf (1935).

The period between the two wars was defined (Michea 1964) as the 'heroic era' of the frequency dictionaries. Several corpus-based projects aimed at producing frequency dictionaries for language teaching were carried out in a number of countries after World War I. Furthermore, the Prague linguistic school, in the same period, provided the theoretical foundations for the study of the quantitative aspects of the language, in the structural linguistic paradigm (Trubetzkoy 1939).

In the 1950s and 1960s, following the introduction of computers to large national lexicographical projects, their use spread to various sectors of the humanities, resulting in the production of indexes, concordances, and frequency counts. This work constituted the nucleus of a new disciplinary field called 'literary and linguistic computing' (LLC), a name that first appeared in 1966. As the volume of textual data available for statistical studies increased, researchers began to focus on the methodological problems facing such studies. Some (for instance Heilmann 1963) claimed that the probability of occurrences of a linguistic unit in a text, estimated on the basis of corpus frequency, was a pertinent feature of the unit. The theoretical interpretation of the frequency in a corpus was discussed within the framework of the relation between a sample corpus and the population (the language as a whole) (Moreau 1962). Large textual corpora formed the basis for experimental and evaluative work carried out on the use of statistical techniques as tools for stylistic studies, as had been proposed in the 1950s and early 1960s, in particular by the French school.

In the early 1960s, the fields of machine translation and of corpus-based statistical linguistics displayed a keen awareness of each other on various occasions, for instance at the 1961 Besançon Congress[1] on 'La Mécanisation des recherches lexicologiques', organized by B. Quemada, and at the first International Congress of AILA (Association internationale des langues appliquées) in 1964 in Nancy, where the two major foci were machine translation and frequency dictionaries for language teaching. Unfortunately, in the years that followed computational linguistics virtually ignored the corpus-based quantitative approach to the study of language; this was particularly true in the Anglo-Saxon countries, where the influence of the generative-transformational paradigm were stronger. The debate between Chomsky and Herdan (1964) on the significance of statistical data and methods in linguistics is typical of the period. Corpus-based activities continued in other fields, including, for example, academic lexicography, descriptive linguistics (where corpus research facilitated the collection of evidence of the use of real languages and sublanguages), and, later, commercial lexicography. In the 1960s and 1970s only a few

[1] The Proceedings are to be found in *Cahiers de lexicologie*, 3.

scholars affirmed the need for those working on the automatization of language processing to combine the CL rule-based approach with corpus studies.

The creation of monolingual and bilingual lexical and terminological collections was important to machine-translation projects of the 1950s and early 1960s. Research was directed towards creating specialized hardware for storing large lexica, improving accessing techniques, and studying inflectional and derivational morphology and other related topics. The ALPAC Report (1966), which abruptly halted the majority of machine-translation projects, recommended the promotion of basic research into linguistics and NLP, and, in particular, the development of large-scale grammars and lexica based on the evidence of large textual corpora. In spite of this recommendation, which in the opinion of several researchers marked the first public recognition of CL as an autonomous disciplinary field, CL activities almost completely neglected the creation of lexica, which in practice, until a few years ago, remained restricted to small toy lexica of some dozens of words. One reason was that the development of CL has always been strongly linked to that of contemporary linguistic theories. In the 1960s and 1970s this relationship was essentially one of dependency of CL on linguistics, and linguists' efforts and interests were mainly concentrated on the study of syntax. In the late 1970s the specific needs and requirements of CL began to influence the developments of some linguistic schools in the generative-transformational paradigm, and certainly contributed to the way in which the lexicon became the focus of increasing interest in linguistics (Pollard and Sag 1987). In the last decade, linguistic generalizations have increasingly revealed a lexical dimension, and the study of the lexicon has now acquired in the generative-transformational schools the central role that it always had in the European linguistic tradition.

Long before this, however, practical use was being made of corpus material. In the United States Thorndike and Lorge (1944) presented data on word frequency and range of use; this research later informed the word-list selection of the Thorndike-Barnhardt dictionaries. In the emerging field that came to be known as corpus linguistics, Kucera and Francis constructed the Brown corpus, which was completed in 1964 (Kucera and Francis 1967). Since then this corpus—particularly recently in its tagged form (Francis and Kucera 1982)—has been used as a yardstick for many different types of operations in computational linguistics.

In the UK Michael West's General Service List of English Words (1953), a parallel work to that of Thorndike and Lorge, was widely used by workers in the field of English as a foreign language (EFL), and formed the basis of the word-list of the first learners' dictionary of English, the *Oxford Advanced Learner's Dictionary of Current English* (*OALD*; Hornby 1942). Early in the 1960s Randolph Quirk and his colleagues at the University of London had established the Survey of English Usage (now computerized; held at the Department of English, University College, London) as a lexical resource for those interested in language in use, and this contributed to—amongst many other works—a comprehensive account of English grammar (Quirk *et al*. 1972, 1985) and another EFL dictionary, the *Longman Dictionary of Contemporary English* (*LDOCE*; Procter *et al*. 1978). Large commercial institutions and to a lesser extent universities have for a number of years been acquiring corpus materials for their own use: for instance, the bilingual (English and French) Canadian Hansard corpus, containing excerpts from the proceedings of the Canadian Parliament, is used by a number of research groups in the USA and elsewhere.

The Lancaster-Oslo-Bergen (LOB) corpus was completed in 1978, with the objective of replicating in British English the design and composition of the American Brown corpus (Johansson 1980) and much work has been done also on the tagging of this body of text (Garside *et al*. 1987). In many centres throughout the English-speaking world work has begun on the International Corpus of English (Greenbaum: ongoing), still at the design stage, but intended to replicate the constitution of the Brown and LOB corpora for the major varieties of world English. Corpus-based research into English is also carried out in other European centres, notably in the Netherlands at Nijmegen and Amsterdam (Oostdijk 1987; Aarts and Meijs 1986).

In languages other than English corpus research also flourishes, although there is space here to mention only a few centres where this work is being done. In France the literary corpus underlying the *Trésor de la langue française* dictionary (*TLF*) was built by Bernard Quemada and his colleagues, and is now available to scholars both on-line and (partially) on CD-ROM (Quemada 1983). Researchers from the Romanska Institutionen Department in Stockholm University built the corpus of modern best-selling French novels (cf. Engwall, in this volume). In Denmark the DANwORD corpus was created in the University of Copenhagen (Maegaard and Ruus 1987), and Bergenholz and his colleagues at the Aarhus School of Business produced the one-million-words-per-year corpora DK87–DK91. At the Istituto di Linguistica Computazionale (ILC) in Pisa work continues on the Italian corpus referred to in the chapter in this volume by Calzolari and Picchi. It is beyond the scope of this paper to go into further detail on existing corpora, but it should be noted that currently corpus material is known to exist for research purposes in at least fifteen European languages other than English.[2]

4. COMPUTERS AND LEXICOGRAPHY

The flourishing tradition of European 'academic' lexicography,[3] dating back at least to the dictionary of the Accademia della Crusca, published in 1612, has always maintained that the production of historical dictionaries, dialectal and regional surveys, and scholarly lexica should be based on the manual analysis of large sets of textual data.

The major European academic lexicographical projects adopted the methods and techniques that used computers to produce indexes, build concordances, and do frequency counts ('dépouillements électroniques des textes') as soon as they became available in the early 1950s, as a result of the first experiments of R. Busa in Gallarate (Italy), in a lexical analysis of the works of St Thomas Aquinas (Busa 1951).

We have already mentioned the innovative and seminal work of Bernard Quemada and his colleagues, which gave rise to the impressive TLF dictionary, still to be completed (Quemada 1983). Further European work in corpus-building resulted in the Gothenburg corpus of Swedish (Språkdata 1988), and the Cobuild corpus of modern English, part of the Birmingham Collection of English Text (Sinclair 1987). Both of these corpora formed a resource for lexicographers, the resultant dictionaries being the *Svensk Ordbok* (Allén *et al.* 1987) and the *Collins Cobuild Dictionary* (Sinclair *et al.* 1987).

Another use of computers in lexicography is reflected in the work being done within publishing houses. The van Dale series of bilingual dictionaries, discussed in Al (1983), were constructed on a computational basis. Work done in Denmark is recorded in the DANLEX reports (DANLEX Group 1987). *LDOCE* (see above), published in 1978, was one of the first dictionaries to be compiled on computer, and to hold in its electronic form data which did not appear in the print version. The computerization of *The Oxford English*

[2] These languages are: Catalan, Danish, Dutch, Finnish, French, German, Greek, Hungarian, Italian, Norwegian, Portuguese, Russian, Serbo-Croat, Spanish, and Swedish. There is also an Arabic corpus.

[3] By the term 'academic' or 'institutional' lexicography we designate the various lexicographical activities promoted by public institutions, such as research institutes, language academies, universities, and so on, as distinct from 'commercial lexicography', typically supported by publishing houses. Of course, several borderline cases exist, depending on differing organizational structures in various countries.

Dictionary is documented in 'The history of the Oxford English Dictionary' (*OED*, i. pp. l–lvi). This particular project caught the public interest and was widely reported in the general press (e.g. Burgess 1989; Shenker 1989).

Finally, in the field of computers and lexicography, electronic access to dictionary data must be mentioned. Many researchers are now working on dictionary material obtained in electronic form direct from publishers. Dictionaries thus in use are too numerous to list in full, but include monolingual collegiate dictionaries in many languages, English monolingual dictionaries for foreign learners, and bilingual dictionaries (in the case of the last two types, the explicit statement of linguistic facts is of great help to the user of on-line dictionaries). Some of the more commonly used machine-readable dictionaries (MRDs) of English are the *Webster's Seventh* and *Ninth New Collegiate Dictionaries*, (Gove 1969; Mish 1986), the *Collins English Dictionary* (Hanks 1986), and *The Oxford English Dictionary* 2nd edition (1989); and (among EFL dictionaries) *OALD* (Hornby 1974), *LDOCE* (Procter *et al.* 1978), and *Cobuild* (Sinclair *et al.* 1987). Other MRDs in common use include bilinguals from publishers such as the British Collins, the Dutch Van Dale, and others; and a number of monolinguals in various European languages, including works from Hachette, Robert, Garzanti, and Zingarelli.

For the general public, the advent of versions of diskettes and CD-ROMs allows dictionary consultation on line, even from a PC. Some examples are the *American Heritage Dictionary* in the USA; *The Oxford English Dictionary* in the UK; in France, the *Grand Robert*; and the mixed bag on the twelve-dictionary CD-ROM published in 1989 by the Multilingual Dictionaries Database Group.

5. AVAILABILITY OF RESOURCES

Research into language in use, and hence all lexical research, received a great impetus with the advent of phototypesetting, a process which involved the creation of an electronic version of each text to be printed. These 'printers' tapes', although often difficult to read, furnished the corpus-builder with a new and rich source of electronic text, once agreement had been reached on the intellectual property rights involved. Whereas early corpus-builders had often relied on simple keyboarding of existing printed texts or on the optical scanner for the conversion of print to electronic medium, it now became possible for those with computational expertise and adequate resources to process printers' tapes as corpus material. Scanning and keyboarding techniques continue to be used, of course, particularly in the creation of specialized small corpora where computational expertise and time are limited and it is often easier to scan or rekey a work than to struggle to read the messy printer's tape.

However, book publishing is only part of the story. With the new technology, many other types of electronic textual data became available for research purposes. As a result of the computerization of reporting techniques—from newspaper production through to the recording of parliamentary proceedings—researchers found themselves able to acquire quite large corpora (100 million words and more), and some of the research focus shifted on to the lexical tools needed to extract facts from such large quantities of data. Researchers on both sides of the Atlantic worked on such corpus-handling tools as part-of-speech taggers and tree builders (Church 1988; DeRose 1988; Garside 1987) and parsers (Garside and Leech 1987; Marcus 1980; Hindle 1983*a*, 1983*b*, 1989; Abney 1990; among others). Others began, with the aid of statisticians, to develop routines in which a variety of standard statistical tests are applied to corpora, leading to the identification of different kinds of linguistic phenomena that are not immediately apparent to the human researcher (Church and Hanks 1990; Church *et al.* 1991, and this volume; Hindle 1990; Tzoukermann and Merialdo 1989; Brown *et al.* 1988; Smadja 1991; Smadja and McKeown 1989; Zernik 1989*b*; Brent 1991*a*, 1991*b*). Techniques also began to be devised for handling bilingual and multilingual parallel corpora (Catizone *et al.* 1989; Warwick *et al.* 1990; Klavans and Tzoukermann 1990*a*, 1990*b*).

Again, where book production is concerned: while many of the newly available printers' tapes contained simple running text (e.g. novels, newspapers, plays, poems), a significant minority proved to hold highly systematized material of immediate relevance to those engaged in research into the lexicon. These were, of course, the dictionary tapes. The advent of the machine-readable dictionary added a new and powerful dimension to lexical research. Computational linguists found themselves in possession of a resource which allowed them to focus as never before upon the lexicon.

The first task facing researchers was the conversion of the dictionary sources, which often took the form of a typesetter's tape, into a more machine-tractable form. Although some of the earliest work focused on the conversion of a particular dictionary (Boguraev and Briscoe 1989*a*), more recently there have been efforts to develop general dictionary parsing techniques that can be applied to more than one dictionary (Kazman 1986; Neff and Boguraev 1989). Once the dictionary sources were cleaned up and parsed, the next step was to structure the contents into some type of database form (Byrd *et al.* 1987; Calzolari 1984; Ahlswede *et al.* 1986). At this point the material in the dictionary was ready to be exploited in various ways.

The recognition that the contents of the dictionaries could be enhanced by using a variety of computational techniques was particularly important. That is, researchers quickly became aware that substantial lexical information was implicit in these dictionaries—the most trivial example was the presence of taxonomies implicit in the genus and differentiae structure of definitions—and much of it could be extracted automatically or semi-automatically. MRDs were quickly used to build various types of semantic hierarchies, and efforts then turned to the extraction of a wide variety of lexical relations between words (Ahlswede and Evens 1988; Amsler 1980; Calzolari 1984; Chodorow *et al.* 1985; Markowitz *et al.* 1986; Nakamura and Nagao 1988). This information turned out to be particularly useful for information retrieval (Fox *et al.* 1987; Lesk 1988; Wang *et al.* 1985). Other efforts focused on the grammatical codes present in learners' dictionaries, using them as a basis to construct a lexicon for a parser (Boguraev and Briscoe 1989*b*). The exploitation of implicit information (Atkins *et al.* 1986) also led to the development of a variety of tools and methodologies for accessing the material in MRDs (Byrd *et al.* 1987; Wilks *et al.* 1990). As more and more material was extracted there was a growing interest in creating lexical databases and lexical knowledge bases that built on material available in these dictionaries (Ahlswede *et al.* 1986; Boguraev *et al.* 1989; Boguraev 1993; Picchi *et al.* 1988; Fox *et al.* 1986; Klavans *et al.* 1990; Pin-Ngern *et al.* 1990); some researchers have even suggested building lexical databases that are the result of merging several dictionaries (e.g. Ide and Veronis 1990; but see Atkins and Levin 1991; and Atkins 1993).

Among the research groups that have been particularly active in this field are—in the USA—the Lexical Systems Group at IBM, Wilks and his colleagues at New Mexico State University, and Evens and her colleagues at the Illinois Institute of Technology; and in Europe the ILC in Pisa, the Department of Computational Linguistics at the University of Cambridge, and the Universities of Nijmegen and Amsterdam in the Netherlands. In Japan the Electronic Dictionary Research Institute was established in 1986, with a budget of fourteen billion yen, covering the period up to the fiscal year 1994, with the objective of constructing large-scale flexible general-purpose dictionaries to be used in artificial intelligence and machine translation applications (JEDR 1990). This project involves the creation of a variety of dictionaries, including a concept dictionary, language-specific dictionaries for English and Japanese, and bilingual dictionaries relating the two languages. As this brief survey makes clear, the number of researchers who have access to MRDs is now growing so fast that it is impossible even to try to do justice to all those who are working in this area.

6. REUSABILITY OF RESOURCES

The use of printers' tapes is an excellent example of the reusability of lexical resources, particularly in the case of dictionary texts. Dictionaries are notoriously labour-intensive and costly to produce, and it is satisfying that the material should serve the interests not only of the human dictionary-user but also of user-computers, albeit in an elementary way as yet. Considerable time, effort, expertise, and funds were soon being devoted to cleaning up printers' tapes for use in corpora, to parsing dictionary texts in order to create lexical databases (LDBs), and to enhancing these LDBs by means of sophisticated computational techniques, even including some fairly crude dictionary mapping, so as to make them into lexical knowledge bases (LKBs) (see references in section 5). The research community, aware of this massive expenditure of resources, began to look for ways in which the lexical material produced could be shared amongst all scholars who needed them.

In the mid-1970s various agencies and institutions began to assemble collections of electronic texts for the purpose of literary and linguistic research: a well-known example is the Oxford Text Archive, a library of texts of various languages and periods, created in 1976 by Lou Burnard, Susan Hockey, and their colleagues at the Oxford University Computing Service. A more ambitious step towards sharing resources, however, came in the USA with the Data Collection Initiative of the Association of Computational Linguistics, the work of Donald Walker (whose concept of the ecology of language has been influential in this area), Mark Liberman, and other colleagues (ACL 1989; Liberman 1989). The first CDs, each containing several million words of copyright-free general American English, are now available at very low cost to researchers in academic institutions. (A related project is the construction of a Tree Bank—a database of text parsed into constituent structure trees—by Mitch Marcus and colleagues at the University of Pennsylvania (Marcus *et al.* 1990; Santorini 1990; Brill *et al.* 1990; Marcus and Santorini forthcoming). A similar text collection initiative is under way for British English, in the form of the British National Corpus, being constructed with the help of government funding by a consortium of industrial and academic partners, led by Oxford University Press.[4] The three-year project, which began in January 1991, aims to make available as a research resource a general British English corpus of 100 million words, hopefully including ten million words of transcribed spoken text, together with some basic text-handling tools.

National and international funding agencies are also becoming aware of the urgent need for reusable lexical resources, both textual corpora and databases. In 1989 and 1990 the Speech and Language Technology (SALT) Club in the UK held workshops to allow the whole research community to discuss and prioritize its needs, and the resultant document provided the government research funding agencies with an indication of where support was most needed, and what type of resources would be of greatest benefit to national research (Leech 1990). In the USA a parallel meeting, the Open Lexical and Textual Resources Workshop, was held in 1990, to allow two major US funding agencies to listen to the research community discussing its immediate and long-term needs (Liberman 1990). In Europe a Workshop on Textual and Lexicographic Corpora was held in January 1990 under the auspices of the Commission of the European Communities (CEC) resulting in recommendations to the Commission on the need for national funding agencies in Europe to support the building of such corpora as shareable, precompetitive resources. Further workshops are planned. The European Commission is also active in sponsoring research aimed at encouraging the reusability of lexical resources, and projects such as Acquilex (on polytheoretical issues, co-ordinated by ILC, Pisa, Italy), Multilex (on standards, co-ordinated by CAP-Gemini and Triumph-Adler), and Eurotra-7 (on reusability of lexical and terminological resources, co-ordinated by University of Stuttgart Institut für maschinelle Sprachverarbeitung) should be mentioned in this regard. The Survey of Language Data in Machine-Readable Form was begun in 1990 by D. Walker and A. Zampolli, with the same aim. This project was initially undertaken in co-operation with INK International, Amsterdam, and detailed questionnaires were sent out to over 600 individuals and institutions world-wide. It was sponsored by many prestigious international organizations, including for instance the European Science Foundation (ESF), the Directorate General XIII of the CEC, and academic associations such as the Association for Computational Linguistics (ACL), Association for Computing and the Humanities (ACH), Association of Literary and Linguistic Computing (ALLC), European Association for Lexicography (Eura-

[4] The other consortium partners are Longman Publishers, Chambers Publishers, the Universities of Lancaster and Oxford, and the British Library.

lex), and many others. The results are available in the form of an electronic database.

Recognizing that overcoming the 'lexical bottleneck' was a common goal facing researchers in computational linguistics, Roy Byrd of IBM proposed the organization of a Consortium for Lexical Research (CLR) (Byrd 1989*a*, n.d.). The CLR would serve as a repository of 'precompetitive' lexical research resources maintained for its participants, who would include industrial, government, and educational organizations. Participants would contribute to and withdraw from this repository. The hope is that the availability of shared resources will facilitate and promote research relating to the lexicon. The Consortium has since been established by the ACL and has recently been funded by DARPA (the Advanced Research Projects Agency of the US Department of Defense). The University of New Mexico is the host to the Consortium, with Yorick Wilks as Director.

Another instance of efforts aimed at the sharing of lexical resources is to be found in the Text Encoding Initiative (TEI), a project begun in 1987 with the purpose of developing guidelines for the preparation and exchange of machine-readable texts for scholarly research, which would satisfy a broad range of uses by the language industries. Work is currently in progress on creating sets of tags for marking features of texts, coded in the framework provided by the Standard Generalized Markup Language (SGML) (see Johansson, this volume); the first draft recommendations have appeared (Sperberg-McQueen and Burnard 1990). The TEI (the brainchild of Nancy Ide) is jointly supported by the ACH, the ACL, and the ALLC; it has also received funding from the US National Endowment for the Humanities, the CEC, and the Andrew W. Mellon Foundation.

7. DISCIPLINARY CROSS-FERTILIZATION: CONFERENCES, COLLABORATION

Computational lexicology and lexicography are, as it were, crossbreeds amongst the scholarly disciplines, and workers in these fields need more than the knowledge and skills of a single discipline. As a result, the 1980s continued an already flourishing tradition and brought together—in conferences, symposia, institutes and summer schools—researchers in various subjects with a common interest in a computational approach to the lexicon: theoretical linguists, computational linguists, computer scientists, and lexicographers, among others.

In 1983 a meeting at SRI International offered researchers with common interests in lexical data, particularly MRDs, an environment in which these interests could be discussed and professional links forged. As early as 1981 a similar service had been offered to European scholars by the European Commission in Luxembourg, which sponsored a conference entitled 'Lexicography in the Electronic Age'. In 1986 there was a seminal symposium of international researchers from a wide range of disciplines at Grosseto in Italy; this symposium, entitled 'Automating the Lexicon', was sponsored by the CEC, the University of Pisa, and the ILC, under the auspices of a variety of professional organizations. The papers from this symposium are published in Walker *et al.* (1994). In the same year a more specialized workshop on the lexical entry was held in conjunction with the Linguistic Society of America's Summer Linguistic Institute in New York. These meetings were important in strengthening the interdisciplinary ties that had already begun to characterize this field.

The professional contacts amongst workers in different branches of study with a common interest in computational techniques for lexical research were renewed at the LSA Summer Institute in 1987 in Stanford, where, in parallel with a regular programme on lexical issues, a series of meetings of the so-called 'Pisa group' were held. This working party, which met during the period 1986–7, was funded and sponsored by the ACL and the ILC, and included representatives from all the disciplines already mentioned. Its purpose was to evaluate—through collaborative work on lexical data (a study was made of a number of verb types in English)—the feasibility of a polytheoretical lexical database, accessible to scholars regardless of the linguistic framework in which they worked (Walker *et al.* 1987).

In 1988 the ESF sponsored and funded a Summer School in Computational Lexicology and Lexicography, under the direction of Antonio Zampolli, at the Istituto di Linguistica Computazionale in Pisa, where researchers and teachers made contacts with scholars working in neighbouring disciplines; this volume is one of the tangible results of such contacts. Again in 1988, a Workshop on Lexical Acquisition was organized in Detroit, in conjunction with IJCAI (International Joint Conference on Artificial Intelligence), and brought together the growing group of researchers who were trying to extract lexical knowledge from MRDs and on-line text corpora using a variety of computational techniques (Zernik 1989*a*, 1989*b*, 1991). This workshop was the first meeting entirely devoted to the exploitation of available on-line resources—whether dictionaries or texts—in the construction of lexical databases and lexical knowledge bases.

Towards the end of the 1980s the rising interest in matters of computational lexicology and lexicography was reflected in the number of tutorial sessions dedicated to these at conferences: for example, the tutorial on MRDs at the 1987 Applied ACL Meeting, the 1990 ACL tutorial on corpus tagging, and most of the 1990 tutorials at COLING (Conference on Computational Linguistics). In 1992 there was also a tutorial on large text corpora at the ACL Meeting. Finally, further evidence of this growing interest is to be found in the recent formation of SIGLEX (Special Interest Group on the Lexicon) within the ACL.

All these events provided opportunities for people with a common interest in computers and the lexicon to meet and to make new contacts. Researchers from widely differing fields of study were brought into professional collaboration, and this new cross-disciplinary approach is reflected in the authorship of the papers in this volume.

8. SETTING THIS VOLUME IN CONTEXT

At the present time, then, we see a significant conjunction of circumstances: the large-scale availability of on-line lexical materials as the by-product of the photocomposition process; the development of quite sophisticated lexical tools for the extraction of linguistic data from these materials; the advent of machine-readable dictionaries and the enhancement of their information content by computational techniques; the measures begun by the research community and the funding agencies to build a store of reusable resources open to scholars world-wide, and to promote the concept of 'precompetitiveness' amongst industrial organizations generating and requiring electronic lexical data. The study of the lexicon—which was already becoming more important in the thinking of theorists, computational linguists, and others—has gained great impetus from the explosion of new resources and the development of new tools, allowing the effective processing of amounts of text which, even ten years ago, would have seemed unmanageably large. It would not be an exaggeration to say that linguistic researchers are standing at the threshold of a new era.

This volume is a product of the ESF Pisa Summer School (see section 7 above), which drew together many of the authors, and others with a common interest in computational approaches to the lexicon. The purpose of this book is—as its title suggests—to set out current work in this area of research, and the chapters themselves have been selected, not only to reflect the principal issues involved, but also to represent the points of views of researchers in the various disciplines where the computer forms an integral part of work on the lexicon. As far as corpora are concerned we limit our discussion to the design and construction of text corpora; speech processing and acoustic corpora lie outside the scope of this volume. The chapters fall into three groups, representing three principal aspects of computational lexicology and lexicography, theoretical and applied: these are data acquisition, theoretical infrastructure, and methodology and tools.

Corpus-builders are faced with two initial decisions: selecting the texts to be included, and deciding on the level of mark-up (annotation and structuring) of the material when it is available on-line. Both of these aspects are addressed in this book. Engwall discusses corpus design features and the criteria available to the researcher who sets out to construct a 'representative' corpus. For illustration, she draws on her own experience in designing the Swedish corpus of modern best-selling French novels. In his chapter, reflecting his contribution to the Text Encoding Initiative (see section 6 above), Johansson summarizes what is involved in the basic encoding of electronic textual resources. Johansson aptly points out that the transformation of texts into textual resources is also a process of interpretation and that therefore compilers of on-line text corpora have the responsibility typically associated with an editor. As he goes step by step through the conversion process, Johansson shows the various challenges that are encountered, from the well-known problems posed by the encoding of special characters to the less obvious problems of encoding editorial comments.

Once the corpus has been assembled, the data may be further processed to make it more accessible and informative. Tools for tagging and parsing are important here, as are statistical methods that compute lexically relevant facts on the basis of large amounts of corpus data. Two chapters in this volume address the issues of processing corpora from each of these perspectives. Hindle considers syntagmatic relationships in his discussion of the problems of parsing a corpus tagged using the part-of-speech tagger described in Church (1988). Church *et al*. focus on paradigmatic relationships, and describe in detail the use of statistical methods to extract lexical data from text corpora, using the Associated Press newswire for their examples. Earlier work (Church and Hanks 1990) discusses mutual information—a measure of association—as a useful tool for extracting collocations from lexical data. This chapter proposes another tool—a statistical measure of substitutability—which can be used to select a set of words which are syntactically and lexically 'substitutable', that is, whose members stand in the same statistical relationship to each other with respect to co-occurrence with the same lexical item. As the authors show, an obvious application of this tool is the identification of potential sets of synonyms.

Part III ('The Theoretical Infrastructure of Lexical Analysis') contains several different types of chapters, in an attempt to give a representative cross-section of current work on the theoretical underpinnings of computational lexicology and lexicography. Within linguistics, the emphasis of work in this area has so far been rather on the syntactic than the semantic component of the lexicon, and the selection of chapters in Part III reflects this fact.

First, the syntactic component: Zaenen and Engdahl's chapter confronts the question of what various syntactic theories require in a lexicon, and asks how the results of this work can be of help to lexicographers and other compilers of lexical resources. The chapter approaches this question by considering the characteristic syntactic properties of several types of sentential-complement-taking verbs, and then examining the requirements such verbs impose on a lexicon within two syntactic frameworks: GB and LFG. The authors point out that the information which the two syntactic models need to handle these types of verbs is quite similar. Nevertheless, the chapter ends on a cautionary note: to the extent that our understanding of classes of lexical items is still incomplete, there are large areas where linguistic theory has less to offer the designer of lexical resources than one might wish.

Following this essay are three chapters which may be considered to some extent as case-studies, each within a different theoretical framework. Gross presents aspects of the painstaking and detailed work being done at the Laboratoire d'Automatique Documentaire et Linguistique at the Université de Paris VII on the construction of a lexicon-grammar for French. In particular, he examines some of the difficult issues that arise in recording the properties of verbs, giving detailed examples of how these will be treated. Uniting the syntactic and semantic components, to a certain degree at least, is Hajičová's discussion of types of grammatical data in a lexicon, which reflects the functional sentence perspective of the Prague school. Her focus is on those lexical facts that play a part in determining word order. Dik discusses the requirements placed on the lexicon by a computational functional grammar (FG), drawing his material from the lexical component of a Prolog implementation of an FG of English. The chapter presents a minimalist view of the lexicon: the lexicon contains only that information about lexical items that is not otherwise derivable. Lexical entries provide information concerning form, meaning, and collocational properties; each type of information is surveyed and exemplified in the paper.

A different aspect of the semantic component is considered in the chapter by Pustejovsky and Boguraev, who discuss, from the point of view of knowledge representation, a range of theoretical issues that enter into the construction of this component in the lexicon. They propose that this component should be seen as a 'generative' rather than 'static' system. This view of the lexicon, they argue, will more easily allow the handling of novel and ambiguous uses of words, as it makes explicit certain facets of word meaning that help determine what kinds of productive relationships words may participate in.

Nirenburg's chapter gives a detailed account of the problem faced by the creator of a lexical knowledge-acquisition system that is to be used for machine translation. This work is one part of a wider research programme being undertaken at the Center for Machine Translation at Carnegie-Mellon University.

In the last essay in this section, Fillmore and Atkins describe the product of collaborative work between a theoretical linguist and a professional lexicographer, using data extracted from a text corpus. In response to the lexicographer's problems (arising ironically from too much knowledge about the definiendum), the theorist proposes a different framework within which to describe the meaning of the word, showing how this new perspective sheds some light on some of the lexicographical problems, allowing a subtler and more faithful description of the interaction of semantics and syntax in our daily use of words.

Finally, in Part IV ('Methodology and Tools') Nagao's chapter is a straightforward account of the rationale and methodology applied in the construction of a terminological dictionary—as opposed to one of general language—at Kyoto University: the *Iwanami Encyclopedic Dictionary of Computer Science*. This is a project which has exploited computational resources to the full, both as lexicographical evidence and as tools in a machine-assisted process of compilation. In contrast, general-language lexicography forms the subject of the chapter by Weiner, who, drawing on his experience in the preparation of the second edition in twenty volumes of the *Oxford English Dictionary*, discusses machine-assisted compilation from the point of view of the editor of a large-scale and extremely sophisticated scholarly dictionary, and looks into the future of the lexicographer–computer relationship.

Turning now to the needs of the lexical researcher; the last chapter, by Calzolari and Picchi, describes in considerable detail the design and functioning of a system constructed at the ILC in Pisa. This system goes some way towards providing the resources needed not only by lexicographers but also by scholars using on-line resources. In particular, the authors give an account of a workstation which increases the value of textual data by linking it to a structured machine-readable dictionary database.

Perhaps prophetically, the volume ends on that note. Computational lexicology and lexicography is a young science, and its practitioners face many challenges. Their response to these will shape its future. We believe that the essays in this volume demonstrate that researchers in this new field are already using these challenges creatively. The first steps in the computational approach to the lexicon have been taken, and the stage is set for exciting new developments.

REFERENCES

Aarts, J., and Meijs, W. (1986) (eds.), *Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora*, Amsterdam: Rodopi.

— and van den Heuvel, Theo (1985), 'Computational Tools for the Syntactic Analysis of Corpora', *Linguistics*, 23: 303–35.

Abney, S. (1990), 'Rapid Incremental Parsing with Repair', in *Electronic Text Research: Proceedings of the Sixth Annual Conference of the Centre for the New OED*, Waterloo: University of Waterloo: 1–9.

ACL (1989), 'The ACL Data Collection Initiative', announcement, Association for Computational Linguistics.

Ahlswede, T. E., and Evens, M. (1988), 'Generating a Relational Lexicon from a Machine-Readable Dictionary', *International Journal of Lexicography*, 1: 214–37.

— — Rossi, K., and Markowitz, J. (1986), 'Building a Lexical Database by Parsing Webster's Seventh Collegiate Dictionary', in *Advances in Lexicology: Proceedings of the Second Annual Conference of the Centre for the New OED*, Waterloo: University of Waterloo: 65–78.

Al, B. F. P. (1983), 'Principes d'organisation d'un dictionnaire bilingue', in B. F. P. Al and Spa (eds.), *Le Dictionnaire: Actes du Colloque Franco-Néerlandais, avril, 1981*, Lille: Presses universitaires de Lille.

Allén, S., *et al.* (1987) (eds.), *Svensk Ordbok*, Stockholm: Esselte Studium.

ALPAC Report (1966), *Language and Machine: Computers in Translation and Linguistics*, Washington, DC: National Research Council Automatic Language Processing Advisory Committee.

Amsler, R. A. (1980), 'The Structure of the Merriam-Webster Pocket Dictionary', doctoral dissertation, University of Texas, Austin, Tex.

Atkins, B. T. S. (1993), 'Building a Lexicon: The Contribution of Lexicography', in M. Bates and R. Weischedel (eds.), *Challenges in Natural Language Processing*, Cambridge: Cambridge University Press: 37–75.

— Kegl, J., and Levin, B. (1986), 'Implicit and Explicit Information in Dictionaries', in *Advances in Lexicology: Proceedings of the Second Annual Conference of the Centre for the New OED*, Waterloo: University of Waterloo: 45–63.

— and Levin, B. (1991), 'Admitting Impediments', in Zernik (1991: 233–62).

Berlin, B., and Kay, P. (1969), *Basic Color Terms: Their Universality and Evolution*, Berkeley, Calif.: University of California Press.

Bloomfield, L. (1933), *Language*, New York: Holt.

Boguraev, B. (1993), 'Building a Lexicon: The Contribution of Computers', in M. Bates and R. Weischedel (eds.), *Challenges in Natural Language Processing*, Cambridge: Cambridge University Press: 99–134.

— and Briscoe, T. (1989*a*) (eds.), *Computational Lexicography for Natural Language Processing*, London: Longman.

— — (1989*b*), 'Utilising the LDOCE Grammar Codes', in Boguraev and Briscoe (1989*a*: 85–116).

— Byrd, R. J., Klavans, J. L., and Neff, M. (1989), 'From Structural Analysis of Lexical Resources to Semantics in a Lexical Knowledge Base', position paper prepared for the Workshop on Lexicon Acquisition, IJCAI, Detroit.

Brent, M. (1991*a*), 'Automatic Semantic Classification of Verbs from their Syntactic Contexts: An Implemented Classifier for Stativity', in *Proceedings of the 5th European Meeting of the Association for Computational Linguistics*.

— (1991*b*), 'Automatic Acquisition of Subcategorization Frames from Untagged Free-Text Corpora', in *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*: 209–14.

Bresnan, J. (1982) (ed.), *The Mental Representation of Grammatical Relations*, Cambridge, Mass.: MIT Press.

— and Kanerva, J. (1989), 'Locative Inversion in Chichewa: A Case Study of Factorization in Grammar', *Linguistic Inquiry*, 20: 1–50.

Brill, E., Magerman, D., Marcus, M. P., and Santorini, B. (1990), 'Deducing Linguistic Structure from the Statistics of Large Corpora', in *Proceedings of the DARPA Speech and Natural Language Workshop, June 1990*: 275–82.

Brown, P., Cocke, J., della Pietra, S., della Pietra, V., Jelinek, F., Mercer, R., and Roossin, P. (1988), 'A Statistical Approach to Language Translation', in *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest: 71–4.

Brugman, C., and Lakoff, G. (1988), 'Cognitive Topology and Lexical Networks', in S. Small, G. Cottrell, and M. Tanenhaus (eds.), *Lexical Ambiguity Resolution*, Los Altos, Calif.: Morgan Kaufman: 477–508.

Burgess, Anthony (1989), 'Taking their Time', *The Observer* (Apr.).

Busa, R. (1951), 'Sancti Thomas Aquinatis hymnorum ritualium: Varia specimina concordantiarum', in *Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate*, Milan: Bocca.

Byrd, R. J. (1989*a*), 'Large Scale Cooperation on Large Scale Lexical Acquisition', paper presented at the Workshop on Lexical Acquisition, IJCAI, Detroit.

— (1989*b*), 'Discovering Relationships among Word Senses', in *Dictionaries in the Electronic Age: Proceedings of the Fifth Annual Conference of the Centre for the New OED*, Waterloo: University of Waterloo: 7–79.

— (n.d.), 'A Consortium for Lexical Research', unpublished MS, IBM T. J. Watson Research Center, Yorktown Heights, NY.

— *et al*. (1987), 'Tools and Methods for Computational Lexicology', *Computational Linguistics*, 13: 219–40.

Calzolari, N. (1984), 'Detecting Patterns in a Lexical Database', in *Proceedings of the 22nd Annual Meeting of the Association for Computational Linguistics*, Stanford, Calif.: Association for Computational Linguistics: 170–3.

— (1991), 'Lexical Databases and Textual Corpora: Perspectives of Integration for a Lexical Knowledge-Base', in Zernik (1991: 191–208).

Catizone, R., Russell, G., and Warwick, S. (1989), 'Deriving Translation Data from Bilingual Texts', in Zernik (1989*a*).

Chodorow, M. S., Byrd, R. J., and Heidorn, G. E. (1985), 'Extracting Semantic Hierarchies from a Large On-line Dictionary', in *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago: 299–304.

Chomsky, N. A. (1957), *Syntactic Structures*, The Hague: Mouton.

— (1965), *Aspects of the Theory of Syntax*, Cambridge, Mass.: MIT Press.

— (1970), 'Remarks on Nominalization', in R. Jacobs and P. Rosenbaum (eds.), *Readings in English Transformational Grammar*, Waltham, Mass.: Ginn: 184–221 (reprinted in N. A. Chomsky, 1972: *Studies on Semantics in Generative Grammar*, The Hague: Mouton, 11–61).

— (1981), *Lectures on Government and Binding*, Dordrecht: Foris.

— (1986), *Knowledge of Language: Its Nature, Origin and Use*, New York: Praeger.

Church, K. W. (1988), 'A Stochastic Parts Program and Noun Phrase Parser for Unrestricted text', in *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Tex.

— Gale, W., Hanks, P. W., and Hindle, D. (1991), 'Using Statistics in Lexical Analysis', in Zernik (1991: 115–64).

— and Hanks, P. W. (1990), 'Word Association Norms, Mutual Information and Lexicography', *Computational Linguistics*, 16/1.

DANLEX Group (1987) (E. Hjorth *et al.* eds.), *Descriptive Tools for Electronic Processing of Dictionary Data*, Lexicographica Series Maior 20, Tübingen: Max Niemeyer.

Derose, S. J. (1988), 'Grammatical Category Disambiguation by Statistical Optimization', *Computational Linguistics*, 14: 31–9.

Fillmore, Charles, J. (1968), 'The Case for Case', in E. Bach and R. T. Harms (eds.), *Universals in Linguistic Theory*, New York: Holt, Rinehart & Winston: 1–88.

— and Atkins, B. T. S. (1992), 'Towards a Frame-Based Lexicon: The Semantics of RISK and its Neighbors', in A. Lehrer and E. Kittay (eds.), *Frames, Fields and Contrasts*, Hillsdale, NJ: Lawrence Erlbaum Associates: 75–102.

— Kay, Paul, and O'Connor, Mary Catherine (1988), 'Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone', *Language*, 64: 501–8.

Fox, E. A., *et al.* (1986), 'Building the CODER Lexicon: The Collins English Dictionary and its Adverb Definitions', published as Technical Report 86-23. Blacksburg, Va.: Department of Computer Science, Virginia Tech.

— Nutter, T., Ahlswede, T., Evens, M., and Markowitz, J. (1987), 'Building a Large Thesaurus for Information Retrieval', in *Proceedings of the 2nd ACL Conference on Applied Natural Language Processing*, Austin, Tex.: Association for Computational Linguistics.

Francis, W. N., and Kucera, H. (1982), *Frequency Analysis of English Usage: Lexicon and Grammar*, Boston: Houghton Mifflin Co.

Garside, R. (1987), 'The CLAWS Word-Tagging System', in Garside *et al.* (1987: 30–41).

— and Leech, F. (1987), 'The UCREL Probabilistic Parsing System', in Garside *et al.* (1987: 66–81).

— Leech, G., and Sampson, G. (1987) (eds.), *The Computational Analysis of English*, London: Longman.

Garzanti (1984), *Il nuovo dizionario italiano Garzanti*, Milan: Garzanti.

Gazdar, G., Klein, E., Pullum, G., and Sag, I. (1985), *Generalized Phrase Structure Grammar*, Cambridge, Mass.: Harvard University Press.

Gove, P. B. (1969) (ed.), *Webster's Seventh New Collegiate Dictionary*, Springfield, Mass.: Merriam-Webster.

Greenbaum, S. (ongoing) (ed.), *ICE Newsletter*, London: University College.

Grimshaw, J. (1981), 'Form, Function, and the Language Acquisition Device', in C. L. Baker and J. J. McCarthy (eds.), *The Logical Problem of Language Acquisition*, Cambridge, Mass.: MIT Press: 165–82.

Gruber, J. S. (1965), 'Studies in Lexical Relations', doctoral dissertation, MIT, Cambridge, Mass. (also published in J. S. Gruber (ed.), *Lexical Structures in Syntax and Semantics*, Amsterdam: North-Holland Publishing House, 1976).

Hale, K. L., and Keyser, S. J. (1986), 'Some Transitivity Alternations in English', Lexicon Project Working Paper 7, Cambridge, Mass.: Center for Cognitive Sciences, MIT.

— — (1987), 'A View from the Middle', Lexicon Project Working Paper 10, Cambridge, Mass.: Center for Cognitive Science, MIT.

Halliday, M. K. (1986), 'Lexis as a linguistic level', in C. Bazell, J. Catford, M. K. Halliday and R. H. Robins (eds.), *In Memory of J. R. Firth*, London: Longman.

Hanks, P. W. (1986) (ed.), *Collins English Dictionary* (*CED*), London: Collins Publishers.

Heilmann, L. (1963), 'Considerazioni statistico-matematiche e contenuto semantico', *Quaderni dell'Istituto di Glottologia VII*, Bologna: Universitá di Bologna: 34–45.

Herdan, G. (1964), 'Quantitative Linguistics or Generative Grammar', *Linguistics*, 4: 56–65.

Hindle, D. (1983*a*). 'Deterministic Parsing of Syntactic Non-fluencies', in *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, Morristown, NJ: 123–8.

— (1983*b*), 'User Manual for Fidditch', published as Naval Research Laboratory Technical Memorandum 7590-142.

— (1989), 'Acquiring Disambiguation Rules from Text', in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*: 118–25.

— (1990), 'Noun Classification from Predicate Argument Structures', in *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, Morristown, NJ: 268–75.

Hornby, A. S. (1942) (ed.), *Oxford Advanced Learner's Dictionary of Current English*, 1st edn., Oxford: Oxford University Press.

— (1974) (ed.), *Oxford Advanced Learner's Dictionary of Current English*, 3rd edn., Oxford: Oxford University Press.

Householder, F. W., *et al.* (1964), *Some Classes of Verbs in English*, Bloomington, Ind.: Indiana University Press.

— *et al.* (1965), *More Classes of Verbs in English*, Bloomington, Ind.: Indiana University Linguistics Club.

— and Saporta, S. (1961) (eds.), *Problems in Lexicography* (3rd edn. 1975), Bloomington, Ind.: Indiana University.

Ide, N., and Veronis, J. (1990), 'Mapping Dictionaries: A Spreading Activation Approach', in *Electronic Text Research: Proceedings of the Sixth Annual Conference of the Centre for the New OED*, Waterloo: University of Waterloo: 52–64.

Jackendoff, R. S. (1983), *Semantics and Cognition*, Cambridge, Mass.: MIT Press.

— (1990), *Semantic Structures*, Cambridge, Mass.: MIT Press.

JEDR (1990), 'An Overview of the EDR Electronic Dictionaries', published as JEDR Technical Report 024, Tokyo: Japanese Electronic Dictionary Research Institute.

Johansson, S. (1980), 'The LOB Corpus of British English Texts: Presentation and Comments', *ALLC Journal 1*, Oxford: Oxford University Press.

Katz, J. (1972), *Semantic Theory*, New York: Harper & Row.

— and Fodor, J. A. (1963), 'The Structure of a Semantic Theory', *Language*, 39: 170–210, repr. in J. A. Fodor and J. Katz (eds.), *The Structure of Language*, Englewood Cliffs, NJ: Prentice-Hall: 479–518.

— and Postal, P. (1964), *An Integrated Theory of Linguistic Descriptions*, Cambridge, Mass.: MIT Press.

Kazman, R. (1986), 'Structuring the Text of the Oxford English Dictionary through Finite State Transduction', published as Technical Report TR-86-20, Waterloo: University of Waterloo.

Klavans, J. L., Chodorow, M. S., and Wacholder, N. (1990), 'From Dictionary to Knowledge Base via Taxonomy', in *Electronic Text Research: Proceedings of the Sixth Annual Conference of the Centre for the New OED*, Waterloo: University of Waterloo: 110–32.

— and Tzoukermann, E. (1990*a*), 'Linking Bilingual Corpora and Machine-Readable Dictionaries with the BICORD System', in *Electronic Text Research: Proceedings of the Sixth Annual Conference of the Centre for the New OED*, Waterloo: University of Waterloo: 19–30.

— — (1990*b*), 'The BICORD System: Combining Lexical Information from Bilingual Corpora and Machine-Readable Dictionaries', in *Proceedings of the 14th International Conference on Computational Linguistics*, Helsinki.

Kucera, H., and Francis, W. N. (1967), *Computational Analysis of Present-Day American English*, Providence: Brown University Press.

Lakoff, G. (1971), 'On Generative Semantics', in D. Steinberg and L. Jakobovits (eds.), *Semantics*, Cambridge: Cambridge University Press: 232–96.

— (1987), *Women, Fire, and Dangerous Things*, Chicago: University of Chicago Press.

Lakoff, G. and Johnson, M. (1980), *Metaphors We Live By*, Chicago: University of Chicago Press.

Laughren, M., and Nash, D. (1983), 'Warlpiri Dictionary Project: Aims, Method, Organization and Problems of Definition', *Australian Aboriginal Lexicography: Papers in Australian Linguistics*, 15: 109–33.

Leech, G. (1990) (ed.), *Proceedings of a Workshop on Corpus Resources, Oxford, January, 1990*, London: DTI Speech and Language Technology Club.

Lees, R. B. (1960), *The Grammar of English Nominalizations*, The Hague: Mouton.

Lesk, M. (1988), 'Can Machine-Readable Dictionaries Replace a Thesaurus for Searches in Online Catalogs?', in *The Uses of Large Text Databases: Proceedings of Third Annual Conference of the Centre for the New OED*, Waterloo: University of Waterloo: 65–74.

Levin, B. (1993), *English Verb Classes and Alternations: A Preliminary Investigation*, Chicago: University of Chicago Press.

Liberman, M. (1989), 'Text on Tap: The ACL/DCI', *Proceedings of the DARPA Speech and Natural Language Workshop, October 1989* (San Mateo, Calif.: Morgan Kaufman).

— (1990), 'Open Lexical and Textual Resources', report from the Workshop on Open Lexical and Textual Resources, Philadelphia: University of Pennsylvania.

McCawley, J. D. (1973), *Grammar and Meaning*, Tokyo: Taishukan.

— (1979), *Adverbs, Vowels, and Other Objects of Wonder*, Chicago: University of Chicago Press.

Maegaard, B., and Ruus, H. (1987), 'The Compilation and Use of a Text Corpus', in A. Cappelli, L. Cignoni, and C. Peters (eds.), *Studies in Honour of Roberto Busa S.J.*, Pisa: Giardini editori: 103–21.

Marcus, M. P. (1980), *A Theory of Syntactic Recognition for Natural Language*, Cambridge, Mass.: MIT Press.

— and Santorini, B. (forthcoming), 'Building very Large Natural Language Corpora: The Penn Treebank', *Computational Linguistics*.

— — and Magerman, D. (1990), 'First Steps towards an Annotated Database of American English', published as Technical Report MS-CIS-90-46, Philadelphia: Department of Computer and Information Science, University of Pennsylvania.

Markowitz, J., Ahlswede, T., and Evens, M. (1986), 'Semantically Significant Patterns in Dictionary Definitions', in *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*: 112–19.

Mel'čuk, I. A. (1973), 'Lexical Functions in Lexicographic Descriptions', in *Proceedings of the Berkeley Linguistics Society 8*, Berkeley, Calif.: University of California: 427–44.

— (1988), 'Semantic Description of Lexical Units in an Explanatory Combinatorial Dictionary: Basic Principles and Heuristic Criteria', *International Journal of Lexicography*, 1: 165–88.

— Iordanskaja, L. N., and Arbatchewsky-Jumarie, N. (1981), 'Un nouveau type de dictionnaire: le Dictionnaire explicatif et combinatoire du français contemporain', *Cahiers de lexicologie*, 38: 3–34.

Michea, R. (1964), 'Les Vocabulaires fondamentaux', in *Recherche et techniques nouvelles au service de l'enseignement des langues vivantes*, Strasbourg: Université de Strasbourg: 21–36.

Miller, G. A. (1990) (ed.), *International Journal of Lexicography*, 3/4, Oxford: Oxford University Press.

— and Fellbaum, C. (1991), 'Semantic Networks of English', *Cognition*, 41: 197–229.

— — Kegl, J., and Miller, K. (1988), 'WORDNET: An Electronic Lexical Reference System Based on Theories of Lexical Memory', *Revue québécoise de linguistique*, 17: 181–211.

— and Johnson-Laird, P. N. (1976), *Language and Perception*, Cambridge, Mass.: Harvard University Press.

Mish, F. C. (1986) (ed.), *Webster's Ninth New Collegiate Dictionary*, Springfield, Mass.: Merriam-Webster.

Moreau, R. (1962), 'Au sujet de l'utilisation de la notion de fréquence en linguistique', *Cahiers de lexicologie*, 3: 150–8.

Nakamura, J., and Nagao, M. (1988), 'Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation', in *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest.

Neff, M., and Boguraev, B. (1989), 'Dictionaries, Dictionary Grammars and Dictionary Entry Parsing', in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*.

Oostdijk, N. (1987) (ed.), *TOSCA: The Nijmegen Research Group for Corpus Linguistics*, Nijmegen: University of Nijmegen.

Pesetsky, D. (1982), 'Paths and Categories', doctoral dissertation, MIT, Cambridge, Mass.

Picchi, E., Peters, C., and Calzolari, N. (1988), 'Implementing a Bilingual Lexical Database System', in T. Magay and J. Zigany (eds.), *BudaLex '88 Proceedings*, Budapest: Akadémiai Kiadó: 317–29.

Pin-Ngern, S., Evens, M., and Ahlswede, T. (1990), 'Generating a Lexical Database for Adverbs', in *Electronic Text Research: Proceedings of the Sixth Annual Conference of the Centre for the New OED*, Waterloo: University of Waterloo: 95–109.

Pollard, C., and Sag, I. (1987), *Information-Based Syntax and Semantics*, i: *Fundamentals*, Stanford, Calif.: Center for the Study of Language and Information, Stanford University.

Procter, P., *et al.* (1978) (eds.), *Longman Dictionary of Contemporary English* (*LDOCE*), London: Longman.

Quemada, B. (1983), 'Les Bases de données informatisées et les dictionnaires', in B. F. P. Al and Spa (eds.), *Le Dictionnaire: Actes du Colloque Franco-Néelandais, avril, 1981*, Lille: Presses universitaires de Lille.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1972), *A Grammar of Contemporary English*, London: Longman.

— — — — (1985), *A Comprehensive Grammar of the English Language*, London: Longman.

Rappaport, M., and Levin, B. (1988), 'What to Do with Theta-Roles', in W.

Wilkins (ed.), *Syntax and Semantics 21: Thematic Relations*, New York: Academic Press: 7–36.

Rappaport, M., Levin, B., and Laughren, M. (1988), 'Niveaux de représentation lexicale', *Lexique*, 7: 13–32 (in English as 'Levels of Lexical Representation', in J. Pustejovsky (ed.), *Semantics and the Lexicon*, Dordrecht: Kluwer).

Roeper, T., and Siegel, M. (1978), 'A Lexical Transformation for Verbal Compounds', *Linguistic Inquiry*, 9: 199–260.

Rosch, E. (1975), 'Cognitive Representations of Semantic Categories', *Journal of Experimental Psychology: General*, 104: 192–233.

— and Lloyd, B. B. (1978) (eds.), *Cognition and Categorization*, Hillsdale, NJ: Lawrence Erlbaum Associates.

— and Mervis, C. (1975), 'Family Resemblances: Studies in the Internal Structure of Categories', *Cognitive Psychology*, 7: 573–605.

Ross, J. R. (1972), 'Act', in D. Davidson and G. Harman (eds.), *Semantics of Natural Language*, Dordrecht: Reidel: 70–126.

Santorini, B. (1990), 'Part-of-Speech Tagging Guidelines for the Penn Treebank Project', published as Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania, Philadelphia.

Shenker, Israel (1989), 'The Dictionary Factory', *New Yorker* (3 Apr.).

Sinclair, J. M. (1987) (ed.), *Looking Up*, London: Collins Publishers.

— *et al.* (1987) (eds.), *Cobuild Dictionary of the English Language*, London: Collins Publishers.

Smadja, F. (1991), 'Macrocoding the Lexicon with Co-occurrence Knowledge', in Zernik (1991: 166–90).

— and McKeown, K. (1989), 'Automatically Extracting and Representing Collocations for Language Generation', in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*: 252–9.

Somers, H. L. (1987), *Valency and Case in Computational Linguistics*, Edinburgh: Edinburgh University Press.

Sperberg-McQueen, C. M., and Burnard, L. (1990) (eds.), *Guidelines for the Encoding of Machine-Readable Texts for Interchange*, Chicago: ACL-ACH-ALLC Text Encoding Initiative.

Språkdata (1988), *Studies in Computer-Aided Lexicology, for Sture Allén*, Gothenburg: Department of Computational Linguistics, University of Gothenburg.

Starosta, S. (1988), *The Case for Lexicase*, London: Pinter.

Stowell, T. (1981), 'Origins of Phrase Structure', doctoral dissertation, MIT, Cambridge, Mass.

Thorndike, E. W., and Lorge, I. (1944), *The Teacher's Wordbook of 30,000 Words*, Columbia, NY: Teachers' College.

Trubetzkoy, N. S. (1939), 'Grundzuge der Phonologie', *Travaux du Cercle Linguistique de Prague*, 7.

Tzoukermann, E., and Merialdo, B. (1989), 'Some Statistical Approaches for Tagging Unrestricted Text', unpublished MS, IBM T. J. Watson Research Center, Yorktown Heights, NY.

Vossen, P., Meijs, W., and den Broeder, M. (1989), 'Meaning and Structure in Dictionary Definitions', in Boguraev and Briscoe (1989*a*: 171–92).

Walker, D., Zampolli, A., and Calzolari, N. (1987) (eds.), 'Towards a Polytheoretical Lexical Database', working paper, Istituto di linguistica computazionale, CNR, Pisa.

—— —— —— (1994) (eds.), *Automating the Lexicon*, Oxford: Oxford University Press.

Wang, Y.-C., Vandendorpe, J., and Evens, M. (1985), 'Relational Thesauri in Information Retrieval', *Journal of the American Society for Information Science*, 36: 15–27.

Warwick, S., Hajic, J., and Russell, G. (1990), 'Searching on Tagged Corpora: Linguistically Motivated Concordance Analysis', in *Electronic Text Research: Proceedings of the Sixth Annual Conference of the Centre for the New OED*, Waterloo: University of Waterloo: 10–18.

Wasow, T. (1977), 'Transformations and the Lexicon', in P. Culicover, A. Akmajian, and T. Wasow (eds.), *Formal Syntax*, New York: Academic Press: 327–60.

—— (1985), 'Postscript', in P. Sells (ed.), *Lectures on Contemporary Syntactic Theories*, Stanford, Calif.: Center for the Study of Language and Information, Stanford University: 193–205.

Wilks, Y., Fass, D., Guo, C.-M., McDonald, J. E., Plate, T., and Slator, B. M. (1989), 'A Tractable Machine Dictionary as a Resource for Computational Semantics', in Boguraev and Briscoe (1989*a*: 192–228).

—— —— —— —— —— —— (1990), 'Providing Machine Tractable Dictionary Tools', *Machine Translation*, 5: 99–154.

Zernik, Y. (1989*a*) (ed.), *Proceedings of the First International Lexical Acquisition Workshop, IJCAI 11*, Detroit.

—— (1989*b*), 'Lexical Acquisition: Learning from Corpus by Capitalizing on Lexical Categories', in *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit.

—— (1991) (ed.), *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Zingarelli, N. (1971), *Vocabolario della lingua italiana*, Bologna: Zanichelli.

Zipf, G. K. (1935), *The Psycho-biology of Language: An Introduction to Dynamic Biology*, Cambridge, Mass.: MIT Press.