1

Introduction

DONALD E. WALKER, ANTONIO ZAMPOLLI,
AND NICOLETTA CALZOLARI

This volume developed out of a Workshop entitled 'Automating the Lexicon: Research and Practice in a Multilingual Environment' that was held in Marina di Grosseto, Italy, 19–23 May 1986. The Workshop was organized in response to a dramatic increase in interest in the lexicon during the early 1980s that was expressed by people and organizations representing a spectrum of activities across many dimensions. However, it is not appropriate to consider the contents as just another workshop proceedings. Instead, the papers have provided a baseline and a reference point for further research and development on problems associated with the lexicon. In addition, from the papers and the discussions a set of recommendations has emerged that has actually served to organize research in the field, further co-operation, and maximize the effectiveness of individual and group efforts.

This Introduction is divided into two sections. The first section deals directly with the Workshop itself, its papers, and the recommendations. The second section describes some of the consequences of the Workshop and identifies subsequent related developments in the lexical field.

In hindsight, after six and a half years, it is clear that the Workshop at Marina di Grosseto marked a turning-point in the field and can therefore be considered as a point of departure for the major events taking place since then. It was precisely this consideration which convinced us to publish the book now, together with the volume which resulted from a Summer School Course organized at Pisa in 1988 (Atkins and Zampolli 1994). Although this book has 'historical' relevance, it addresses many issues that are still being debated today and that guide research and development efforts.

1. The Workshop Perspective

1.1 Interest in the lexicon as a background to the Workshop

The interest in the lexicon that motivated the Workshop can be attributed to a number of factors, all of which are still relevant:

1. Theoretical developments within linguistics are placing increasing emphasis on the lexical component. It is proving to be a central source of semantic as well as syntactic information.

2. Demonstrations of the feasibility of applications of natural language processing are creating demands for large-scale systems in industry and in national and supranational organizations. For these systems to be practical they must deal with tens and even hundreds of thousands of lexical items.

3. The effort required to create comprehensive dictionaries for these purposes is substantial. It may prove to be the most costly and time-consuming task in such developments. Currently, each system is building its own lexicon, and there is increasing recognition that the duplication of effort is enormously expensive. However, differences in the organization and content of these lexicons make it difficult or impossible to share linguistically relevant information across systems.

4. The computational linguistics community is becoming increasingly conscious of the extensive resources contained in published dictionaries, and explorations are under way to determine how that information in machine-readable form can be exploited to expedite system development.

5. Publishers are beginning to realize the potential of their dictionaries for commercial purposes. They are recognizing the value of establishing lexical data bases from which a variety of dictionaries can be derived. They are also becoming aware of the breakdown of the distinction between different reference works (dictionaries, fact books, encyclopaedias).

6. Increased communication among lexicologists, lexicographers, linguists, computational linguists, publishers, and commercial natural language processing software developers has led to a heightened awareness of common objectives and the complementarity of skills and knowledge.

7. Initial experiments have given support to the idea that it may be possible to construct 'neutral lexicons' that can be shared, with different theories selecting relevant linguistic information through an appropriate interface.

1.2 Organization of the Workshop

The Workshop was sponsored by the Commission of the European Community (DGXIII), the University of Pisa, and the Institute for Computational Linguistics of the Italian National Research Council. It was also held under the auspices of the Council of Europe, the European Science Foundation, the Association for Computational Linguistics, the Association for Literary and Linguistic Computing, Euralex, and the Commission on Computational Linguistics of the International Association of Applied Linguistics. The Organizing Committee was composed of Nicoletta Calzolari, Loll Rolling (CEC, DGXIII, Luxembourg), Juan Sager (University of Manchester Institute of Science and Technology, Manchester), Don Walker, and Antonio Zampolli (chair). The second in a series, it built on its predecessor, which was organized by Don Walker and Robert Amsler and held in April 1983 at SRI International in Menlo Park, California.

Where the first workshop examined the machine-readable dictionary from the perspective of the research community, the publishers, and the emerging market intermediaries, the second was much broader. Its purpose was to explore research efforts, current practice, and potential developments in work on the lexicon, machine-readable dictionaries, and lexical knowledge bases with special consideration for the problems created by working with different languages. The intent was to identify the current state of affairs and to recommend directions for future activities.

To help in the realization of these objectives, a set of papers was solicited for the Workshop under the following general headings: Research Areas, Core Problems, Application Areas, and Developing Research Resources. People were asked to prepare comprehensive surveys and evaluations of activities going on in the field. We also requested reports on national projects in related areas. At the end of the agenda a 'Consolidation' session was scheduled to consider the following topics: the lexical entry as a basis for integration, co-operation and communication, priorities for research and development, and the next steps.

The participants were chosen to bring together representatives from the different kinds of areas that we believed were relevant to the various problems associated with the lexicon. This concern led us to invite linguists, lexicographers, lexicologists, computational linguists, artificial intelligence specialists, cognitive scientists, psycholinguists, publishers, lexical software marketers, translators, funding agency representatives, and professional society representatives. We wanted this heterogeneous mix of people to get to know each other and to explore their similarities and differences in interests.

The Workshop was organized to further discussion. At the meeting, a discussant actually presented the paper, although the author was the first to comment. This departure from customary practice was intended to stimulate an examination of the topics under consideration by having the material presented concisely and from two perspectives at once: that of the author and that of the discussant. Our hope was that such an arrangement would result in our spending more time examining the ideas in the papers, rather than the specifics of their contents. The results confirmed this expectation, and it was a lively Workshop indeed. Actually, even in those cases where the discussant was less 'constructive', the spirit of the group carried through quite well.

The papers that were prepared for the Workshop do identify the state of affairs at that time better than any other material available then or since. They also, in the context of their discussion at the Workshop, led to a series of recommendations for future work. The papers and these recommendations will be considered in the next two sections.

1.3 The papers

The papers that are actually included in this volume are a subset of the papers prepared for and distributed in advance of the Workshop. They contain the material most relevant for and influential in affecting subsequent developments in the lexical area. Most of them were circulated widely in pre-print form, which accounts for the pervasiveness of their effect. The majority of them have been updated since their original preparation, so a substantial part of the material presented here is actually new.

As for the other papers prepared for the Workshop, several were published elsewhere and so could not be included. The rest of the papers either addressed issues from a non-computational standpoint, focused on technological aspects pertinent at that time and not updated by their authors, or described the status of developments in various countries that are no longer of current interest.

1.3.1 *Research areas*

Linguistics: In 'Identifying the Linguistic Foundations for Lexical Research and Dictionary Design' Richard Hudson examines the positions taken by 'mainstream' linguistics on the lexicon, contrasting them with his own approach, 'Word Grammar'. He finds that the two kinds of approaches have quite different consequences for the lexicographer.

Semantics: In 'Approaches to Lexical Semantic Representation' Beth Levin reviews the types of evidence used to argue for lexical semantic representations of verb argument structures and evaluates a range of approaches to lexical semantic representation that are found in the lexical semantics literature.

Parsing: In 'Lexical Information for Parsing Systems: Points of Convergence and Divergence' Robert Ingria presents the kinds of lexical information needed for parsing, considers how such information is expressed in existing parsing systems, and explores the possibility for sharing information among the lexicons of different systems.

Generation: In 'The Lexicon in Text Generation' Susanna Cumming reviews the kinds of lexical information that have been included in generation systems and how they interact with other parts of those systems, and then examines the range of co-occurrence phenomena and the implications they have for optimal lexicon design.

1.3.2 *Application areas*

Office practice: In 'Dictionary Systems for Office Practice' Roy Byrd examines the use of computerized dictionaries by both people and programs in the office environment, presenting a viewpoint on how dictionary systems should be constructed and what functions they should provide to the different classes of users.

Translation: In 'The Role of Dictionaries and Machine-Readable Lexicons in Translation' Jonathan Slocum and Martha Morgan consider the differential use of lexical data by human translators and machine translation programs, concluding that bilingual dictionaries are most likely to be relevant for both classes of users.

Education: In 'Machine-Readable Dictionaries and Education' Judy Kegl examines in detail the problems posed in education by the use of current dictionaries, some current research, and the prospects for educational applications of electronic versions of dictionaries properly augmented and with multiple access paths. This paper was actually prepared shortly after the Workshop, although much of the material presented here was discussed there.

Information retrieval: In 'Why Use Words to Label Ideas: The Uses of Dictionaries and Thesauri in Information Retrieval' Michael Lesk shows that thesaurus relations are central to current information retrieval systems, pointing out that while question-answering systems might find that dictionaries provide more refined distinctions, it would be at the risk of missing related concepts.

1.3.3 *Developing research resources*

Using machine-readable dictionaries: In 'Machine-Readable Dictionaries and Research in Computational Linguistics' Branimir Boguraev examines the broad range of problems in computational linguistics to which machine-readable dictionaries have been applied, and argues that far different structures and organizations will be required for more effective and reliable utilization in the future.

Structure and access: In 'Structure and Access in an Automated Lexicon and Related Issues' Nicoletta Calzolari describes the advantages of moving from the static view of traditional machine-readable dictionaries to lexical data bases which provide a dynamic structure that allows a much broader variety of functions and relations to be explored and exploited.

1.3.4 *International activities*

Europe: In 'Automated Lexical Resources in Europe: A Survey' Susan Armstrong-Warwick provides a comprehensive review of the various types of computerized dictionaries currently available or under development in Europe, concentrating on languages other than English.

In section 2.2 we will review some of the changes that have taken place in work on the lexicon following the Marina di Grosseto Workshop.

1.4 The recommendations

As important as the papers were in establishing a baseline for the current state of research and practice in the field at that time, the recommendations we arrived at as a result of our deliberations are proving to be equally significant for contemporary developments. The Workshop has clearly identified, among a range of academic and industrial research and development groups, publishers, and commercial firms that market lexical products, a convergence of interests that would motivate the establishment of large computational lexical resources intended for shared activities. As will be apparent, the complementarity of monolingual and multilingual concerns was consistently stressed.

1. Create and maintain registries of machine-readable dictionaries and related resources, lexical data bases, text corpora, bibliographic references and the corresponding documents, and human resources: where appropriate, establish designated repositories for materials that are available for distribution.

2. Establish terminological conventions for working with lexical resources that can be shared by groups working with computers as well as those using more traditional approaches.

3. Clarify the copyright issues associated with the various lexical resources and establish a framework that supports the broadest distribution of these materials to groups of relevant users.

4. Organize a lexical data entry group with responsibilities for

identifying lexical materials that should exist in machine-readable form, for determining a standard format or set of formats in which they should be represented, and for arranging to have them coded and made available through the repositories.

5. Establish a communication network, progressively computerized, that can link together the Workshop participants and other interested people and groups to allow sharing information about new and continuing developments and to provide a forum for examining critical issues.

6. Establish more general communicatión channels through professional societies, their journals and newsletters, and presentations at regular conferences (e.g. the Association for Computational Linguistics (ACL), Euralex, the International Association for Applied Linguistics (AILA), the Association for Literary and Linguistic Computing (ALLC), and the Association for Computers and the Humanities (ACH)).

7. Arrange special meetings that promote further communication, co-ordination, and co-operation for both general and specialized interest groups focused on selected topics.

8. Study the work of lexicographers to model their behaviour, incorporating the results in knowledge-based systems that support lexicographic activities.

9. Study how people interact with standard and electronic dictionaries and lexical data bases to determine the most effective procedures for human/machine interaction.

10. Develop lexical and lexicographic workstations embodying resources, data, and tools that directly support lexicological and lexicographic activities.

11. Investigate new technologies and products that could be incorporated into such workstations; correspondingly, identify design characteristics that would facilitate working with lexical materials and try to motivate their development as products.

12. Support 'internship' and 'sabbatical' links that allow people in various disciplines to work closely with each other on project activities.

13. Develop curricula, courses, texts, and manuals for lexicology and lexicography that will further interdisciplinary understanding and that can be used in a variety of education and training contexts.

14. Compare and contrast lexical information, particularly in the form of 'lexical entries', as reflected in logical and linguistic theories, computational linguistic systems, machine-readable dictionaries, translation activities, and lexicographic practice in order to identify dimensions of similarities and differences; based on those dimensions, create a metaformat that subsumes the structures of the various types of information to be included, and that can be used both as a reference

frame for evaluation and exchanges and as a model of a computerized meta-lexicon from which lexicons for different research and applications may be derived.

15. Establish procedures for converting the contents of machine-readable dictionaries, text corpora, and other lexical data into formats appropriate for a range of computational needs.

16. Apply frequency measures to gather systematic and representative synchronic and diachronic data on a broad range of language variables in text corpora.

17. Determine whether dictionaries can be designed so that they can be used in both human and machine environments.

18. Convince publishers to begin saving the photocomposition tapes of books, journals, and other published materials and to make them available for research.

19. Establish project designs and patterns of co-operation that promote sharing of data, tools, and human resources (particularly scarce ones) among academic and industrial research and development groups, publishers, and commercial firms that market lexical products.

20. Create linguistic data bases from existing and newly produced sources that embody machine-readable dictionaries and large text corpora (thousands of millions of words), and create tools that make it possible to explore their relationships systematically.

21. Create lexical data bases and explore their utility for supporting the creation of general and specialized dictionaries, monolingual and bilingual dictionaries, encoding and decoding dictionaries.

22. Establish procedures for deriving monolingual and bilingual lexical and lexicographic material from text corpora: of particular interest are strategies for identifying phrases, synonyms, hyponyms, and other classes of relationships automatically.

23. Establish large collections of 'evaluated' paired and aggregated translations reflecting bilingual and multilingual sources, and develop procedures for exploring and exploiting their correspondences.

24. Develop methodologies for interrelating monolingual and bilingual dictionaries: explore the possibility of combining technical monolingual dictionaries with bilingual general dictionaries to create technical bilingual dictionaries.

25. Establish lexical indices for determining and representing stylistic features, subject-matter codes, and other sociolinguistic parameters: create procedures for incorporating them in machine-readable dictionaries and for using them for lexicological and lexicographic research.

26. Study the use of lexical information by children and conduct experiments to determine what kinds of lexical resources would be most effective for educational purposes.

27. Establish a range of dictionaries that reflect needs for specialized information or non-standard modes of interaction (e.g. handicapped users), clarifying their similarities and differences as well as the feasibility of deriving them from standard dictionaries or from each other.

28. Develop new programming languages that support the co-ordinated manipulation of strings (text sequences) and structures (taxonomies, frames, and logical relationships).

29. Develop new data base designs that allow storing, accessing, and interrelating (at detailed feature levels) both the form and content of multibillion word text files.

30. Study pictures, tables, diagrams, and other illustrative material and develop workstation tools for processing them and for relating them systematically to corresponding machine-readable dictionary entries and passages in text corpora.

31. Develop a theory of pictorial reference that will facilitate relating lexical and semantic information to images.

32. Extend and modify the typology of traditional dictionaries and lexical tools and resources so that it applies to materials that are now or will be in machine-readable form and to the emerging uses of these materials.

33. Determine how to incorporate the experiences of traditional academic, humanistic, and classical studies for lexical and lexicographic research.

After providing in section 2.3 a description of projects and major activities that have taken place in the lexical area since the Marina di Grosseto Workshop, in section 2.4 we identify the way in which subsequent work has addressed these recommendations.

2. Developments since 1986

2.1 Initial steps taken

Because of their historical interest, it is appropriate to identify a number of developments in the field that were motivated directly by the Marina di Grosseto Workshop.

The most immediate event was a workshop on 'The Lexical Entry', which took place at the City University of New York in conjunction with the Summer Linguistic Institute in July 1986. It was convened to concentrate on the nature and structure of lexical entries, following up specifically on the Marina di Grosseto discussions. A major objective was determining whether it is possible to establish a comprehensive framework for viewing the lexicon that can be shared by the disparate elements in the community. We examined in detail how different theoretical frameworks and system implementations influence the format for a lexical entry. One goal was to characterize a general representation or 'metaformat' that would subsume the specific ones. We realized that not everyone would agree to share a single model. Rather, we hoped to identify the range of parameters that are being used and from which different approaches could be viewed as making a selection. Significant progress was made, although these issues are still being debated today.

Some other immediate developments were a panel on 'The Lexicon in a Multi-lingual Environment', held at COLING '86 in Bonn, August 1986; the establishment of an *ad hoc* working group on 'Computational Lexicology and Lexicography' by the European Science Foundation; the formation of a specialist working group on 'Dictionaries and the Computer' at Euralex in Zurich, September 1986; a conference on 'Standardization in Lexicography' in Saarbrücken, October 1986, supported by the European Science Foundation; a conference on 'Advances in Lexicology' at the Centre for the New OED, in Waterloo (Canada), November 1986; a session on 'Words and World Representations' at TINLAP3, Theoretical Issues in Natural Language Processing, in Las Cruces (New Mexico), January 1987; a small working group, the Pisa Polytheoretical Group, exploring the possibility of a neutral lexicon in Pisa in March 1987 and at a number of other meetings in the following years; a workshop on 'The Lexicon in Theoretical and Computational Perspective' during the Summer Linguistic Institute at Stanford (California), July 1987; a special double issue on the lexicon in the journal *Computational Linguistics* in 1987; and a summer school on 'Computational Lexicography and Lexicology' held in Pisa during July and August 1988.

Beyond 1988, the developments in the field occurred in such profusion that it would be difficult to catalogue them all. In addition, it is particularly appropriate to stop with the Pisa Summer School, since the material presented there forms the basis for the second volume in this series of publications.

2.2 Some developments in work on the lexicon since the Workshop

Perhaps the major development in theoretical and computational linguistics with regard to the lexicon is the central role that the lexicon has assumed in many frameworks. Two in particular have achieved a greater prominence in the computational field: Categorial Grammar and Lexicalized Tree Adjoining Grammar (TAG). In both these formalisms, there is no grammar outside the lexicon. In Categorial Grammar, individual lexical items specify the constituents they combine with and the constituents that result. In Lexicalized TAG, individual lexical entries specify fragments of tree structure ('basic trees') associated with terminal elements. In both cases, lexical entries effectively combine both subcategorization and phrase structure information, thereby eliminating the need for independent phrase structure rules.

Work in lexical semantics and argument structure has continued to receive substantial attention within linguistics and within the related fields of psycholinguistics, computational linguistics, and lexicography. Interest in this area is evidenced by a string of conferences, workshops, and tutorials on lexical semantic issues, the formation of a special interest group on the lexicon within the ACL, and the publication of various edited volumes.

Within lexical semantics itself, there have been several new areas of interest. Increased attention is being paid to systematicities involving multiple meanings of words. Certain aspects of meanings have been shown to have an important place in the mapping between lexical semantics and syntax. One is the lexical aspectual characterization of verbs; another is the stage/individual level distinction.

The lexicon continues to hold an important place in a variety of linguistic theories. Particularly important, within the last few years, has been the development of lexical mapping theory within Lexical Functional Grammar and the interface between lexical semantics and syntax. The relative importance of meaning and syntax in child language acquisition has been a topic of debate in the child language acquisition literature.

There has been a trend within some work in the Government-Binding framework over the last five or six years to give more syntactic accounts of problems that involve word meaning—through the use of elaborated syntactic structures that are different from surface syntactic representations or through the introduction of empty predicates—that in some ways is reminiscent of generative semantics, although the approach is constrained in rather different ways.

Complex feature-based grammar formalisms are now essentially state of the art. Use of such formalisms and, in particular, the development of typed feature structures and inheritance and default mechanisms have allowed lexicon developers to create more highly structured lexicons and to reduce the amount of information necessary in individual lexical entries.

Most existing theories of lexical representation assume a view of lexical items as collections, perhaps structured, of essentially static word senses. Recently, generative theories of the lexicon have begun to be developed. Under such theories, a core set of word senses, typically with greater internal structure than is assumed in previous theories, is used to generate a larger set of word senses when individual lexical items are combined with others in phrases and clauses.

One of the most interesting proposals for the architecture of such a generative theory of the lexicon would include multiple levels of representation for the different types of lexical information needed. Among such levels are Argument Structure (for the representation of 'arity' information for functional elements, such as verbs and function nouns), Event Structure (for the representation of information related to verbal tense, aspect, and event type), Qualia Structure (for the representation of the defining attributes of an object, such as its constituent parts, purpose and function, mode of creation, etc.), and Inheritance Structure (for the representation of the relation between the lexical item and others in the lexicon).

In recent years, textual corpora, both tagged and untagged, have come to play a major role as training data for natural language processing systems. The availability and use of such corpora have had a number of effects. Large tagged corpora have allowed parsing systems to acquire part-of-speech information automatically. They permit the creation of stochastic part-of-speech taggers that attempt to assign parts of speech even to unknown words, reducing the size of lexicons: for certain applications, only words which are relevant to the application domain need to be present in the lexicon; others can be handled by the tagger. Large corpora, tagged or untagged, allow for the acquisition of other lexical information, including lexical semantic information.

Previously, it was assumed that the lexicon of the system contained lexical entries for all the words appearing in the domain with the complete syntactic and semantic information necessary for the application task. Use of partial and fallback parsing techniques allows the jettisoning of this closed vocabulary assumption. Lexicons of systems using such techniques will certainly contain fairly complete entries, including both syntactic and semantic information, for those words relevant for the application task. Depending on what parsing technique is used and on whether a tagger is available, non-relevant words may also be included, but with more fragmentary information.

Lexical selection has taken centre stage (along with discourse structure concerns) as one of the major theoretical issues in text generation. As the distance between the input to a text generation system and its output grows larger, the lexical selection problem can no longer be simply handled by a one-to-one mapping of 'concepts' to lexical items; more and more systems are making meaningful choices between larger and larger pools of candidate items. Collocation phenomena are being taken more seriously; metaphor is being treated more systematically; the effects of 'pragmatics' in its various senses (including text planning, conversational implicature, and social factors) are being seen as central.

Two separate—and yet very related—developments can be observed in the use of machine-readable dictionaries. On the one hand, both the theoretical linguistics and the computational linguistics communities have become much more aware of the tremendous potential that they hold for large-scale lexicon extraction, for better understanding the nature of the lexicon, and for verification of linguistic hypotheses and theories. On the other hand, return to empiricism in the field of natural language processing has led to an increased activity in the area of corpus-based language studies, and, in particular, corpus-driven acquisition of lexical information.

In addition to pursuing its central goal—semi-automatic derivation of facts about words, word meanings, and word uses—computational lexicology has elaborated a range of other equally important questions. For instance, developing methods and techniques for structuring and analysis of on-line lexical resources, looking for clues to the structure and organization of the human lexicon, and defining methodologies for lexical semantics research are only some of the concerns of the field. An especially pertinent issue, at the core of computational analysis of language on the basis of information available in dictionaries, concerns the relationship between natural language processing, formal syntax and lexical semantic theories, and the way in which this relationship is reflected in the kind of information sought in dictionaries for incorporation into a computational lexicon.

To a large extent, recent work reflects a change in view concerning the predominant paradigm of computational lexicology: whereas early efforts for utilizing dictionary data were aimed primarily at what had been explicitly stated in the dictionary entries, more recent developments have focused on carrying out much more detailed, and global, analysis of the sources with a view of uncovering information which turns out to be systematically encoded across the entire source(s). Thus, examples of extraction processes in the earlier framework include: acquisition of information about control and logical type of predicates, extraction of semantic features (e.g. selectional restrictions) for lexical disambiguation, or derivation of information about stress assignment. In contrast, the desire to make maximal use of the information in dictionaries has promoted work exploiting the distributed nature of the lexical knowledge encoded in these sources; representative examples here include: developing better models of speech recognizer front ends, building semantically sound lexical hierarchies, sprouting networks of lexical relations between words, deriving empirical evidence for the existence of semantically coherent word clusters, refining such networks to reflect word-sense distinctions, and even extending such activities beyond the boundaries of a single dictionary. Populating richer lexical structures introduces an additional dimension to the notion of lexical relation and accounts for the permeability among word senses. In general, there is an observable shift of emphasis, from extracting primarily syntactic properties of words, to seeking and formalizing lexical semantic information.

Finally, recent research has tended to view critically the notion of building a computational lexicon on the basis of existing machine-readable dictionaries; a particularly common feature is a certain amount of scepticism towards attempts to instantiate such a lexicon entirely by automatic means. Still, there is a shared attitude that while there are many ways in which dictionaries might fail as sources, there are also ways to maximize the value of information found in them.

2.3 Projects and major activities in the lexical area: A general perspective

A considerable number of projects and activities in the lexical area have flourished since the Marina di Grosseto Workshop, the conditions and time being ripe for the speeding-up of a major effort in lexical development. These developments took place not only in Europe, but also in the United States and in Japan. Some of the initiatives were, in fact, a direct consequence of having brought together, at the Workshop, the right people at the right moment.

Before enumerating some of those projects and activities, it is appropriate to identify a more general consequence of the Workshop itself. That was the emergence of a perspective that has inspired many of the recent developments and that helps in understanding and placing in context many of the activities that have resulted. The main stress in the last few years has been on the notion of 'reusability', both in the sense of reusing existing lexical resources (e.g. machine-readable dictionaries) and in the sense of building lexical resources that can be used in many different theoretical and applicational frameworks.

This concept of 'reusability'—directly related to the importance of 'large-scale' linguistic (lexical, textual, grammatical, knowledge) resources—has contributed significantly to the structure of research and development efforts. All the large international projects in this area, on both sides of the Atlantic and in Japan, are motivated by this idea. It is particularly appropriate to view it in relation to the emergence and increasing importance of the notion of 'language industry' more generally.

Immediately after the Marina di Grosseto Workshop, the right atmosphere was created to consider the possibility of converging the efforts of various groups towards the common goal of demonstrating the 'feasibility' of building large reusable lexicons, which needed to be both polytheoretical and multifunctional—with respect to applications and possible users. The establishment of the so-called Pisa Polytheoretical Group, an informal group formed by Zampolli, Walker, and Calzolari to study the concept of 'lexical entry', set things in motion. Now the words 'reusability', 'polytheoretical', 'polyfunctional' are keywords characterizing the actual framework of research and development—even though there is not necessarily complete agreement on how these terms are to be applied.

The notion of 'reusability' is now intended in two main senses:

reusable-1: the feasibility of reusing the wealth of existing resources intended primarily for distribution in printed form, although available in machine-readable form as well (e.g. dictionaries, textual corpora), and extracting from them information which can be used in natural language processing applications, thus transforming them into 'reusable-2 resources';

reusable-2: the feasibility of building large-scale linguistic resources—either by entering new data or with the help of reusable-1 resources—which can be reused in the framework of different theoretical frameworks, for different types of applications, and by different user types, both human and machine.

These two aspects of 'reusability' are in a sense prototypically represented by two European projects supported by the Commission of the European Community (CEC). For 'reusable-1': ACQUILEX (ESPRIT BRA), which is directed towards the acquisition of lexical information from machine-readable dictionaries for natural language processing applications. For 'reusable-2': ET-7 (CEC), which involved studying the feasibility of building large-scale reusable lexical and terminological resources.

After these two 'pioneering' projects in the area, the overall development of international projects and activities—funded by both public and private organizations—soon became so extensive that it is not possible to provide here a detailed description of ongoing activities. However, it is particularly appropriate to mention the following European efforts that have been supported primarily by the Commission of the European Community (CEC) and its ESPRIT and Language Research and Engineering programs (LRE), by Eureka, by the European National Science Foundation, and by the Council of Europe:

- Multilex (ESPRIT), establishing a standard for multilingual lexica;

- Genelex (Eureka), producing generic and application specific lexica according to a unified model;
- NERC, a feasibility study for the creation of a Network of European Reference Corpora (CEC);
- ET-6 and ET-9 (CEC), the study and implementation of a common computational framework for the development of a variety of software tools for NLP;
- various ET-10 projects (CEC) on the characterization and implementation of lexical collocations, on semantic analysis of lexical data extracted from the Cobuild dictionary, on terminology and extralinguistic knowledge, and on the statistical acquisition of lexical knowledge;
- Expert Group for Bilingual Corpus-Based Lexicography (Council of Europe);
- European Corpus Initiative (ECI) for collection and dissemination of corpora for the various European languages;
- ELSNET (European Network of Excellence for Natural Language and Speech) Task Group for Reusable Linguistic Resources (ESPRIT);
- Delis (CEC-LRE) aiming at producing descriptive lexical specifications and tools for corpus-based lexicon building;
- Onomastica (CEC-LRE) for the creation of reusable tools and an inventory of proper names for speech and NLP;
- Studies on Reusability of Grammars (CEC-LRE and GRAAL-Eureka);
- EAGLES (CEC-LRE) aiming at accelerating the provision of common functional specifications for the development of large-scale language resources.

Various initiatives in the United States have been established to provide for collecting and distributing lexical data and spoken and written corpora and formulating sharable lexicons and grammars: the Consortium for Lexical Research (CLR); the Data Collection Initiative (DCI); the Common Lexicon Working Group; and the Linguistic Data Consortium (LDC). They are sponsored by the Association for Computational Linguistics (ACL), the US National Science Foundation (NSF), and the US Defense Research Projects Agency (DARPA). In addition, WordNet, a facility that embodies lexical relationships among words and concepts, has been developed at Princeton. It has been supported by various agencies of the US Department of Defense. In Japan, the major activities have been associated with the Electronic Dictionary Research Institute (EDR), which is supported by the Ministry of Trade and Information and various industrial groups.

Internationally, it is important to mention the Survey of Linguistic Resources (SLR), which has been sponsored by a large number of international scientific associations and governmental agencies to identify existing resources in machine-readable form. The Text Encoding Initiative (TEI) is establishing guidelines for the encoding and interchange of written and spoken texts, covering a broad range of linguistic and humanistic analyses, and of terminological and lexical data; it is sponsored by the US National Endowment for the Humanities (NEH), the Commission of the European Community (CEC), the Mellon Foundation, and the Canadian Research Council. The Center for Electronic Texts in the Humanities (CETH) provides a facility for inventorying textual and lexical data, for collecting and providing access to them, and for motivating research to study how they are used and by what kinds of people; it is supported by Rutgers and Princeton Universities, the NEH, and the Mellon Foundation. The picture is becoming even more fascinating—and promising for future results—with the recent efforts to establish formal co-operation between projects and groups in Europe, the United States, and Japan at both institutional and governmental levels.

In addition to its 'scientific' implications, this large intellectual and economic movement obviously entails 'strategic' considerations. It has become essential to define a general organization and plan for research, development, and co-operation to avoid duplication of efforts and provide for a systematic distribution and sharing of knowledge (see e.g. the Final Report of ET-7).

2.4 How the Workshop recommendations map on to subsequent developments (1986–1992)

We list here, for each of the recommendations formulated in Marina di Grosseto, some of the initiatives which can be seen as effective implementations of it. It would be inappropriate to try to elaborate on all of the specific mappings. What is interesting and important to stress is the fact that this set of recommendations established the framework for research and development in the following years. In addition, there are many projects in universities and industrial laboratories that are addressing these issues directly.

1. Registries: CLR, CETH, ET-7, ELSNET Task Group for Reusable Resources;
2. Terminological conventions: Euralex Working Group;
3. Copyright and availability issues: NERC, CLR, CETH;
4. Lexical elements and standards: Pisa Polytheoretical Group,

TEI, ET-7, NERC, EAGLES, ELSNET Task Group for Reusable Resources;

5. Communication network: CLR, ELSNET, EAGLES, CETH;
6. Professional society involvement: note the large increase in the number of lexical papers at recent ACL, ALLC/ACH, COLING, Euralex, and other conferences, and the co-sponsoring by professional societies of initiatives such as the TEI, International Surveys, common tutorials, workshops, panels, etc.;
7. Special meetings: many specialized workshops and conferences have been held, some sponsored by professional societies, some by governmental agencies, some organized by affinity groups;
8. Studying lexicographers: Pisa Summer School on Computational Lexicology and Lexicography, Euralex Workshop on Computational Lexicography in Tampere, Delis;
9. Computer/human interaction: Acquilex, Delis;

10–11. Workstations, new technologies and products: ET-6, ET-9, Delis, Acquilex, Genelex, Multilex, CLR, EDR, ET-10;

12. Internships and sabbaticals: motivated by the many project activities under way in Europe, the United States, and Japan, and often supported by university and government grant programmes;
13. Curricula and texts: OUP book series, Euralex, Erasmus, Pisa Summer School on Computational Lexicography and Lexicology;
14. Structuring lexical entries: Pisa Polytheoretical Group, Acquilex, ET-7, EAGLES;
15. Computationalizing lexical resources: Acquilex, Multilex, Genelex, ET-10, NERC;
16. Statistical data on text corpora: NERC, ET-10, many ongoing research projects;
17. Human- and machine-readable dictionary formats: Acquilex, Delis, Genelex;
18. Collecting machine-readable resources: CLR, DCI, ECI, LDC, CETH, Acquilex, Delis;
19. Resource sharing: CLR, DCI, ECI, LDC, CETH, NERC, EAGLES, ELSNET, the ESPRIT and LRE frameworks for linguistic projects;
20. Textual and lexical data bases: CLR, DCI, ECI, LDC, CETH, NERC, Genelex;
21. Lexical data bases for dictionary development: Acquilex, Multilex, Genelex, Delis, ET-10;

22. Phrase, synonym, and hyponym identification: Acquilex, WordNet, ET-10, NERC, many individual studies;
23. Paired translations: DCI, ECI, LDC, NERC;
24. Relating monolingual and bilingual dictionaries: Euralex Working Group, Acquilex, ET-10;
25. Lexical indices for style parameters: CETH, ET-10;
26. Lexical resources for education: WordNet and various individual projects;
27. Dictionaries for specialized uses: WordNet and various publishers and individual projects;
28. Programming languages combining string and structure manipulation: ET-6, ET-7, ET-9, Acquilex, Delis;
29. New data base designs for massive text files: NERC, Birmingham English Monitor Corpus, individual projects;
30. Relating images to text: many groups are working on multimedia issues, but they are less directly connected to lexical research at the present, except for individual projects;
31. Relating images to lexical and semantic information: again, while the multimedia groups will certainly be addressing these problems, it will become increasingly important for people working on the lexicon to become involved;
32. Broadening the typology of lexical materials: CLR, NERC, individual projects;
33. Relating humanistic studies and lexical research: CETH, workshops of ACL, ALLC, ACH.

2.5 Conclusions

It is clear that exciting things are happening in the world of the lexicon. The Marina di Grosseto Workshop motivated the preparation of a set of important papers that showed where the field was at that time. It got people from quite different backgrounds together to appreciate that they shared problems and could profitably work together. And it produced a comprehensive set of recommendations for guiding both research and development efforts. These contributions have had an effect far beyond our expectations. However, although the work is well begun, there is still a long—and interesting—way to go!

References

Atkins, B. T. S., and Zampolli, A. (1994) (eds.), *Computational Approaches to the Lexicon*, Oxford: Oxford University Press.