Acquisition of Lexical Knowledge for Natural Language Processing Systems

Proposal for ESPRIT Basic Research Actions

B. Boguraev, *University of Cambridge*
E. Briscoe, *University of Lancaster*
N. Calzolari, *University of Pisa*
A. Cater, *University College Dublin*
W. Meijs, *University of Amsterdam*
A. Zampolli, *Institute for Computational Linguistics, Pisa*

Project Coordinator:
Prof. Antonio Zampolli
Instituto di Linguistica Computazionale C.N.R
Via della Faggiola 32, I-56100 Pisa, Italy

Project Contact Officer:
Dr. Branimir Boguraev
Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street, Cambridge CB2 3QG, UK

Acquisition of Lexical Knowledge for Natural Language Processing Systems

1 Summary of research

This document describes a proposal for an integrated European research project to develop techniques and methodologies for utilising existing machine-readable dictionaries (MRDs) in the construction of lexical components of natural language processing (NLP) systems. The research will draw on and extend current work on extracting data from published MRDs and formalising this data to facilitate the algorithmic processing of language. The main focus of the project will be on extending existing techniques for processing single MRDs in a monolingual (and currently mostly English) context to the extraction of lexical information from multiple MRD sources in a multilingual context with the overall goal of constructing a single integrated multilingual lexical knowledge base.

Machine-readable dictionaries are just one type of lexical resource; there is a wide range of data available electronically in the form of lexicographic databases, terminological databanks and existing lexicons for parsing, generation, machine translation (MT) and office automation systems. However, MRDs have recently been shown to be of particular relevance to work in computational linguistics and natural language processing. Research on MRDs has both computational and linguistic aspects. On the one hand, advanced computational techniques for modifying and accessing textual databases need to be developed, which are specifically geared to the organisation of dictionaries. On the other, the project will draw on and extend theoretical linguistic research on the lexicon and the lexical systems of different languages in order to address the related issues of form, content, scale, acquisition and reusability of lexicons for large scale NLP/MT systems.

The emphasis of our research is on identifying the most general and domain-independent aspects of lexical knowledge and expressing this knowledge in a fashion that will make it reusable by a wide variety of NLP systems. The central, long-term, goal of the research programme is the development of a multilingual lexical knowledge base, rooted in a common conceptual/semantic structure which is linked to, and defines, the individual word senses of the languages covered and which is rich enough to be capable of supporting a 'deep' processing model of language. The lexical knowledge base will contain substantial general vocabulary with associated phonological, morphological, syntactic and semantic/pragmatic information capable of deployment in the lexical component of a wide variety of practical NLP systems. The functionality of such a knowledge base will be evaluated by assessing its capability of supporting prototype monolingual (e.g. query processing) and multilingual (e.g. machine translation) systems.

Contents

2 Introduction

The fundamental goal of research on natural language processing is the automation of the various language processing tasks, such as first language learning, text comprehension, speech synthesis or translation. Knowledge of and about words underlies all these tasks, yet until very recently the lexical components of natural language processing (NLP) systems have by and large been the poor sisters of computational linguistic research. Most extant NLP systems have only illustrative lexicons of, at most, a few hundred words; Whitelock *et al.* (1987:233f), for instance, report that various "state-of-the-art" NLP systems under discussion at a recent workshop had an average lexicon size of 25 words (discounting the Rosetta machine translation system which has a vocabulary of approximately 6000 words). As NLP systems become more sophisticated and potentially able to make the transition from laboratory prototype to workplace, the need for large lexicons, which represent lexical information reliably and precisely enough for automated use, becomes more pressing.

The task of constructing a realistic lexicon for a natural language is formidable, not only because of the absence of a well-articulated theory of what it should contain, but also because of the enormous number of words to be dealt with. The *Oxford English Dictionary* (OED) contains entries representing 250,000 independent words approximately. However, even the OED still does not list many words from more specialised fields. Walker and Amsler (1986) highlight this problem by pointing out the considerable divergence between the vocabulary of the *Merriam-Webster Seventh New Collegiate Dictionary* (W7) and the vocabulary which occurs in news reports on the *New York Times* newswire. It would certainly be impractical (given current levels of funding) and probably unproductive for computational linguists to set about constructing substantial lexicons by hand. There are simply too many words and too many distinct types of knowledge about words potentially relevant to different kinds of NLP system. For this reason, a number of researchers have turned to machine-readable versions of published dictionaries as potential sources of lexical information for use by NLP systems. This development is comparatively recent (see Walker, Zampolli and Calzolari, 1988; Boguraev and Briscoe, 1988a) and has been made feasible by the advent of computer typesetting techniques, which have ensured the availability of machine-readable versions of most published dictionaries.

2.1 Machine-readable dictionaries and NLP systems

There are several advantages and at least one major disadvantage to the use of machine-readable dictionaries (MRDs) as sources of lexical knowledge for NLP systems.[1] Firstly, since there is a considerable tradition behind the production of dictionaries for human consumption, we might hope that they will provide a suitable starting point for defining the contents of a lexicon for machine use. Secondly, since many published dictionaries contain substantial amounts of lexical information, much of the construction work has already been done for the computational linguist. On the other hand, published dictionaries are produced with the human reader in mind and therefore make many inconvenient assump-

[1]Machine-readable dictionaries are not the only instance of existing *lexical resources*. Such resources include a wide range of data available electronically in the form of lexicographic databases, terminological databanks and existing lexicons for parsing, generation, machine translation (MT) and office automation systems. This proposal, however, is primarily concerned with studying the information available in, and extractable from, a number of MRDs because they represent the richest and most structured source of lexical information currently available.

tions from the point of view of processing by machine; for example, the assumption that the user can understand definitions of word senses written in natural languages. Nevertheless, it is generally accepted that MRDs provide the most accessible source of information about general (rather than specialised) vocabulary, which must form the starting point for the development of usable commercial NLP systems whatever their particular domain of application. Furthermore, considerable advances have been made in recent years towards the goal of extracting the information contained in MRDs (see Boguraev and Briscoe, 1988b; and Walker and Zampolli, 1988).

The majority of work with MRDs so far, under the general heading of computational lexicography as applied to the concerns of NLP theory and practice, has been carried out on the basis of single dictionaries (see section 2.3 below). There is, however, strong motivation to extend such work to embrace multiple dictionary sources. Given that different dictionaries by their nature emphasise different aspects of knowledge about words, some dictionaries are more suitable for extracting a particular kind of data than others. Parallel studies of more than one dictionary clearly would allow the derivation of a more detailed, complete and reliable body of lexical data which can be of utility to a wider range of NLP programs. Furthermore, studies of dictionaries across languages are going to reveal regularities which can be exploited in the derivation of transfer lexicons for use in machine translation (MT). Finally, multilingual dictionaries can be used as organisational devices for the process of deriving transfer lexicons (see section 3.1 below).

2.2 Goals of proposed research

This proposal describes a project to carry out a coordinated research programme aimed at further developing techniques and methodologies for the utilisation of existing MRDs in the construction of lexical components of NLP systems. Current work with MRDs is carried out predominantly on the basis of English dictionaries (Boguraev and Briscoe, 1988b, discuss such research at length); a notable exception here is the lexical group at the Institute for Computational Linguistics in Pisa, who has studied a number of Italian sources (Zampolli, 1984; Zampolli *et al.*, 1987). The research project we describe will focus on extending the techniques above from work with single MRDs in a monolingual context to the extraction of lexical information from multiple MRD sources in a multilingual context.

The long-term aim of this programme is to develop a multilingual lexical database (LDB) and an associated lexical knowledge base (LKB). A *lexical database* is defined to be a structured body of lexical data — phonological, morphological, syntactic and semantic/pragmatic — capable of deployment in the lexical component of a wide variety of practical NLP systems, ranging from text-to-speech synthesis to machine translation. A substantial proportion of such information can be found, explicitly or implicitly, in machine-readable dictionary sources: significant part of our work will be on consolidating and extending techniques for making this data available to NLP systems. A *lexical knowledge base* is a richer structure which, in addition to incorporating lexical information at different levels (and, in this case, different languages), also imposes a conceptually based organisation on such data, where individual word senses are defined in terms of conceptual 'primitives' and interlinked in a number of semantically important ways. This makes such a structure capable of not only satisfying the lexical requirements of a wide range of NLP systems, but also of supporting a 'deep' (knowledge-intensive) processing model of language.

Therefore, the emphasis of the proposed research is on identifying the most general and domain-independent aspects of lexical knowledge and expressing this knowledge in a

fashion that will make it reusable by a wide variety of NLP systems.

The concept of *reusability* is central to the development of useful LDBs/LKBs. Firstly, such systems must allow existing lexical resources (and in particular MRDs) to be reusable for different applications by separating the extraction of lexical data (and its representation in an application/theory-independent form) from its use by some particular NLP aplication.[2] Secondly, reusability presupposes some notion of a 'standard interchange format' for lexical data from different MRD sources. Thirdly, reusability rests on the assumption that it is possible to 'compile' a genuinely useful 'general purpose' lexical database on the basis of existing MRD sources.

Reusability underlies the separate stages we identify in the process of converting a number of distinct MRDs into an intregrated multilingual LKB. Initially, each MRD will be 'normalised' and converted into a standard format which allows loading into standardised LDB software. The purpose of the LDB is to provide flexible access to dictionary entries via any of the information contained in the MRD and to allow information from different MRDs to be compared, combined or merged to form new more comprehensive lexical resources. In this way, implicit information concerning generalisations across dictionary entries within MRDs (which is obscured by organisation into an alphabetical list) is made available, the reliability of information in a particular MRD can be checked by comparison with others, information from different sources can be integrated, and links can be established via bilingual MRDs for the transfer of lexical information between monolingual MRDs of different languages. The final stage is the production of an integrated multilingual LKB. This will be developed through the exploitation of the facilities provided by the LDB. The essential prerequisites for conversion of the LDB into a LKB are the identification of a common conceptual structure behind the word senses across the different MRDs (both within and across languages) and the establishment of the links between this structure and the individual word sense definitions in each dictionary.

[2]There are several terms used in the MRD literature intended to express this notion of the reusability of information extracted from MRDs, such as "theory-neutral", "polytheoretical", and "theoretically-uncommitted" representation of this information. In what follows, we use the terms "theory-neutral" or "theory-independent" to denote this concept.

3 Theoretical Background

3.1 NLP Systems Require Lexical Knowledge

All NLP systems use lexical knowledge of some kind; for example, a text-to-speech synthesis system requires knowledge concerning the pronunciation of letters and letter sequences as well as of individual words where they diverge from these more general rules. In addition, it will require knowledge of rhythmic patterns of prominence (stress) in the spoken language. Thus, the English letter sequence *cake* can be pronounced according to simple rules for realising *c* and *k* as the phoneme */k/* and a slightly more complex rule which states that a vowel followed by a consonant and *e* as in *ake* is realised as a long vowel, that is */ei/*. However, no such rules apply in the pronunciation of *yacht* so its pronunciation must be stored separately and such rules blocked in this case. Similarly, in polysyllabic words the system must know which syllable carries greatest stress, for example *VIsion*. Within linguistic theory (at least since Bloomfield), it is widely assumed that the lexicon is the source of idiosyncratic, irregular information and that predictable regularities concerning the phonological, morphological, or syntactic organisation of a language are represented independently as part of that language's general grammatical description. Given this approach, the pronunciation of *yacht* would probably be the only one of the lexical facts about English mentioned above which would be represented in the lexicon.

However, the great majority of published dictionaries (and associated MRDs) represent all of these facts in one way or another. So information of this sort (both regular and irregular) is, in principle, recoverable from such sources. However, much of this information is implicit in dictionaries because the recovery of generalisations and rules of the type outlined above requires examination of the relevant parts of many individual lexical entries throughout a dictionary. Much of the existing computational work with MRDs has involved the conversion of conventionally organised sequential dictionaries into lexical databases capable of supporting this type of analysis (see e.g. Calzolari, 1984a and 1984b).

Broadly speaking, most NLP systems make use of three distinct types of knowledge: linguistic knowledge, knowledge about the domain and knowledge about the application.[3] Along a different dimension, a distinction can be made between general world knowledge and application- and domain-specific knowledge. Linguistic knowledge elaborates the phonological, morphological, and syntactic properties of the linguistic fragment which the system covers. The connection between the linguistic knowledge and the system's knowledge about the world is provided by general lexical semantic/pragmatic knowledge. Closed-class, "grammatical" vocabulary is linked closely to the system's morphological and syntactic knowledge, while open-class "contentful" vocabulary serves as an index into the system's knowledge of the world and thus underlies its consequent ability to perform tasks such as answering questions, acquiring new information and generally performing inferences in the context of this knowledge. Thus, in any particular application, general lexical and world knowledge is supplemented with domain-specific lexical and practical knowledge which allows the system to bridge the gap between the linguistic form of the user's input and its particular 'translation' with respect to the application in hand.

MRDs and linguistic knowledge MRDs are a rich source of both regular, productive and irregular, idiosyncratic lexical knowledge. Not all the linguistic knowledge which NLP systems utilise is lexical but there is a lexical dimension to each type of linguistic knowl-

[3] Unrestricted text-to-speech and some other systems do not have the same notion of domain.

edge. For example, syntactic knowledge concerning the ordering and grouping of words into phrases, clauses and sentences can be represented largely independently of particular words. However, this knowledge cannot be deployed satisfactorily without knowledge of the parts-of-speech of words, the number and type of syntactic complements required by different verbs (and to a lesser extent adjectives and nouns), the agreement properties of particular nouns, and so forth. All of this latter type of information, whether regular or irregular, is extractable from extant MRDs — recent examples of studies within this paradigm include the work of Huttenlocher (1983, 1985) and Shipman and Zue (1982), who have looked at partitioning of large lexicons for the purposes of speech recognition; Carter (1987) and Carter *et al.* (1988), who have used an on-line dictionary to study the phonological structure of English and experiment with different models of lexical access; and Boguraev and Briscoe (1987) and Boguraev *et al.*, who have used the same dictionary to study subcategorisation classes and logical types of verbs and to derive a large computational lexicon for Generalized Phrase Structure Grammar (Gazdar *et al.*, 1985). More work of this nature is discussed in Boguraev and Briscoe (1988b).

MRDs and extra-linguistic knowledge In addition to purely linguistic lexical knowledge, many NLP systems require semantic/pragmatic knowledge of words to function effectively. This is true of most machine translation systems, database querying systems, intelligent information retrieval systems, and so forth. The task of manually coding the lexical semantic/pragmatic knowledge required by a practical (rather than prototype) system of this type would be gargantuan and, no doubt, this explains the small size of many NLP system lexicons. Fortunately, MRDs provide this information, albeit in a form which is not readily accessible to a machine. There has been considerable work on extracting the conceptual hierarchies which lie behind the vocabulary of a language by analysing the definitions of word senses in MRDs and explicit synonym/antonym lists provided in the dictionary entries of some MRDs (Amsler, 1983; Calzolari, 1984; and more recently Fox *et al.*, 1988).

3.2 Lexical Limitations of NLP Systems

Current NLP systems, developed with particular applications or research problems in mind, tend to contain manually coded lexicons which suffer from two broad problems; firstly, their vocabulary size is too limited for serious use and secondly, the lexical entries provided will not generalise to other systems. It is obvious that MRDs provide a potential solution to the problem of vocabulary size. However, it is also the case that they can help with the identification of general lexical knowledge which will be relevant to various application domains.

The problem of limited vocabulary Even though the domain-independent components of NLP systems can, in theory, be carried across to a new application without modification, there is no system which contains a complete (or anywhere near complete) set of words or body of facts, above and beyond those which are necessary for its functioning in a particular context. The fact that current systems have limited, domain-specific, vocabularies not only raises questions concerning vocabulary acquisition when transporting, or adapting, an existing system to a new domain or application, but also has implications for its degree of habitability.

This problem arises because systems representative of the state-of-the-art today invariably have lexical and world (general and domain) knowledge bases *constructed by hand*,

which are minimally sufficient for the particular application under consideration. There is a need to create a general lexical and world knowledge base sufficiently wide in scope and coverage, as well as relatively theory-neutral in structure, to be of genuine use to most potential applications without the need for substantial additions. In order to construct a domain-independent lexical knowledge base which is genuinely transportable between applications, it is necessary to define clearly the content and function of such information and clarify the relationship between general and domain knowledge.

Acquisition of new vocabulary Even though this is one of the fundamental prerequisites for genuine transportability (between different domains and applications) and lexicon reusability (between different language processing tasks), there is still no well understood methodology for acquiring domain knowledge and associated vocabulary or for specifying the relationships between the domain and the general knowledge and between the domain and the general vocabulary. It is clear that the processes of constructing domain-specific lexicons and relating their semantic components to the systems' knowledge bases should be sufficiently general, to apply to more than one pragmatic context, and automated as much as possible, to work from existing lexical resources.

However, (lexical) knowledge bases have mostly been constructed by hand and on demand, and with no reference to work on (semi-)automated concept learning or work on (semi-)automated general knowledge acquisition (such as, e.g. the work by Lenat *et al.*, 1986). One approach to this *knowledge acquisition bottleneck* is that taken in SRI's NanoKLAUS (Haas and Hendrix, 1983) where the system learns facts about a domain by leading the user through a controlled dialogue designed to elicit linguistic and domain properties of specialised words and terms. Even though it is possible to acquire new knowledge structures in such a way, the repetitive, inflexible and lengthy nature of the procedure severely constrains the applicability of such a technique for realistic domains.

Resolving the lexical knowledge acquisition bottleneck One way to approach these problems is to attempt to develop a single lexical knowledge base from which lexicons for specific purposes can be constructed on demand. This resource would need to contain all the information about particular words potentially relevant to any NLP system represented in as theoretically neutral a fashion as possible and would require an exhaustive vocabulary. There are similar needs for such rich repository of lexical knowledge in the multi-lingual environments characteristic of MT work. Clearly, these requirements are ideals and it is unlikely that any such system could ever be constructed from which specialised lexicons could be derived merely by a process of selection and filtering. Nevertheless, significant progress in the direction of such a general lexical knowledge base would greatly reduce the amount of effort involved in the construction of lexicons for specific systems as well as providing a useful resource for investigating various lexicographic properties of language.

Clearly, existing MRDs provide sources which cannot be ignored by any research effort aimed at the construction of this system. Thus the fundamental tenet of this research is that utilising MRDs will help define in gereral terms the (minimal) aspects of information which a lexical knowledge base should contain.

3.3 Current Work with MRDs

A substantial amount of work aimed at extracting information from machine-readable sources of published dictionaries has been undertaken (see Boguraev, 1988; Boguraev and

Briscoe, 1988b). This work can be characterised by a number of common tendencies. Firstly, almost all attempts to analyse a machine-readable dictionary contribute, to some degree, to our understanding of converting an on-line MRD into a lexical database, where the lexical information available implicitly in the dictionary is made (more) explicit in the LDB. Secondly, such work typically aims to construct, from dictionary entries devised for human consumption, lexical entries for computational systems for processing language. Thirdly, most efforts are based on the analysis of a single, and monolingual (usually English) MRD. Finally, these efforts are usually motivated by the particular lexical requirements of specific systems, and are thus necessarily driven by the strengths of MRDs, as well as by the particular interests of the groups involved.

Still, even though significant generalisations concerning tools, techniques and methodologies for MRD utilisation are few and far between, the results so far in extracting a wide range of lexical information from such sources are encouraging. We summarise these results below.

Lexical conceptual taxonomy Prominent work under this general heading includes that of, for instance, Amsler (1980, 1981), who has demonstrated the theoretical feasibility of compiling conceptual taxonomies by extracting generalisations across dictionary definitions; Chodorow *et al.* (1985), who have investigated the automation of such a procedure, and Alshawi's implementation of a prototype system capable of extracting fragments of semantic networks, for incorporation into such larger taxonomic structures, from individual word sense definitions (Alshawi, 1988).

Our use of the term 'taxonomy' is intended to denote a theoretically relatively neutral and uncommitted framework for knowledge representation. A number of arguments have been put forward suggesting a degree of conceptual equivalence between a network representation of a hierarchy (sort taxonomy) and a first order (propositional) system; leaving interpretation issues aside, we can assume that structuring conceptual knowledge into frame-like concepts with slot-like role descriptions, organised in an inheritance hierarchy along generalisation/specialisation axes, will make that knowledge generally accessible to NLP systems with differing approaches to the representation of conceptual knowledge (see e.g. Boguraev, 1987).

Calzolari (1984b, 1988) has argued that a lexical database is essential for the study of a number of semantic relations between word entries, and Fox *et al.* (1988) have derived, from an on-line dictionary, a thesaurus-like structure not unlike a taxonomy in the sense used here. Most work of this kind relies critically on the ease and feasibility of parsing dictionary definitions, typically phrased in natural language; and while no extant system to date has achieved a 100% success rate[4], analytical studies suggest that even more accurate special purpose definitions analysers can be developed utilising the findings of e.g. Meijs *et al.* (1988), who have looked at the relationship between structure and meaning in dictionary definitions, and Markowitz *et al.* (1986), who have identified a number of lexical-semantic relations associated with defining constructions.

Not all conceptual information for NLP systems necessarily resides in a taxonomically structured network; consequently, further (and complementary) research with MRDs is aimed at, for example, extracting selectional restrictions, categorising verbs into semantically important classes (e.g. active and stative), and clustering semantically related word senses (e.g. synonyms) together. The most representative of this kind of research is the

[4]For instance, Alshawi reports 80%, and Calzolari about the same.

Lexical Systems Project at IBM Yorktown Heights (Byrd, Calzolari *et al.*, 1987; Chodorow *et al.*, 1988).

An interesting issue, and one that is particularly relevant in the multilingual environment of (knowledge-based) machine translation, is that of shared semantic features and properties across languages. In section 4 below (*General conceptual lexical knowledge*), we present some arguments in favour of a common conceptual taxonomy underlying both the structures of different MRDs and the lexical organisations of different languages. While virtually no work to date addresses this issue, it is clearly one that needs detailed study because a genuinely integrated multilingual LKB must be based on a common conceptual structure.

MRDs and grammatical frameworks Equally intensive research has been carried out in extracting syntactically relevant data from MRDs. In recognition of the importance assigned to detailed grammatical information associated with individual lexical items, more and more publishers are beginning to include elaborate syntactic (part-of-speech, subcategorisation, valency and so forth) tags in their entries. Dictionaries of particular relevance in this context now include not only the *Longman Dictionary of Contemporary English* (Procter, 1978; henceforth LDOCE) and the *Oxford Advanced Learner's Dictionary* (Hornby, 1980; henceforth OALD); recent editions of e.g. Collins (*The COBUILD English Language Dictionary*; Sinclair, 1987), van Dale (van Dale, 1984) and Wahrig (Wahrig, 1986) offer similar information for English, Dutch, and German, respectively.

Not all of these sources have been exclusively used for the extraction of syntactic information; however, both detailed analytical studies of grammar coding systems and specific projects for utilising such systems demonstrate their value for automatic text analysis. Thus Akkerman's comparative analysis of LDOCE and OALD not only makes explicit the range of linguistic phenomena within the descriptive power of particular coding systems, but also provides a critical assessment to the extent of which these systems can be reliably employed for natural language parsing tasks (Akkerman, 1988). Boguraev and Briscoe (1988c), on the other hand, demonstrate how the detailed syntactic information available (explicitly, as well as implicitly) in an on-line dictionary can be used within a number of different grammatical frameworks, while Boguraev *et al.* (1987) present a methodology for reliably deriving, from this dictionary, a large scale computational lexicon for practical natural language processing. Similar goals have been expressed by, e.g. Atwell and Elliott (1987), Calzolari and Antona (1987), Ingria (1984), and Michiels (1982). An *ad hoc* working group is engaged in a preliminary study of the notion of a theory-neutral lexicon, functioning as a common base for the lexical requirements of parsing systems within different grammatical frameworks (Walker, Zampolli and Calzolari, 1987).

To date, extensive work on syntactic subcategorisation has been undertaken only for English and Italian; studies of deriving subcategorisation information from MRD sources have been carried out predominantly for English. However, a multilingual LKB will require such information for all languages covered. Therefore, there is considerable scope for further parallel work with other languages and also for the transfer of such information from English MRDs to monolingual MRDs of different languages. This latter approach is non-trivial but justifiable on the basis of the particular strength of the LDOCE coding scheme and the large amount of research already undertaken on it. To succeed it will require comparison of the syntactic behaviour of classes of words across the relevant languages and establishment of language-independent criteria for transfer of information. An example here, which illustrates one possible approach to such studies, is the identification

of semantic classes of words with common behaviour (e.g. Levin, 1988).

Orthography and phonology The work on extracting lexical information from MRDs is not confined exclusively to the syntactic/semantic dimension. There is a long tradition in using dictionaries to derive a range of orthographic data, including, for example, word lists and character collocations for spelling correction (Yannakoudakis and Fawthrop, 1983), and hyphenation patterns for word processors (Liang, 1983; Knuth, 1986). In addition, more recent work has addressed the question of the utility of these resources for speech analysis and generation: thus Carter *et al.* (1987, 1988) have used an on-line dictionary as the basis of a study of the phonological structure of English and its implications for lexical access, while Church (1985) has applied similar resources to the problem of stress assignment. Briscoe (1985) lists a number of speech processing projects which make use of on-line MRDs for the generic task of compiling special-purpose word lists transcribed into project-specific phonemic alphabets, incorporating primary and secondary stress assignment and marking of syllable boundaries.

As demonstrated by the work cited above, the extraction of such information from dictionary sources is by no means a trivial task. Thus there is a need to produce an LDB with a phonological representation sufficiently rich to capture the various aspects of the phonological structure of language, and appropriate for the different target vocabularies. This would require research aimed at developing robust techniques for parsing the phonetic fields typically found in existing MRDs and reliably assigning e.g. syllable boundaries, stress markers and other structural descriptions to words and word segments.

MRDs and computational lexicography A complementary line of work has necessarily addressed the problems of mounting MRDs on-line. While not all research groups have focussed primarily on the software issues of providing efficient and flexible access into the electronic sources, the bulk of the research effort has gone on normalising the original source and converting the result into some sort of a database in which entries can be accessed and compared on the basis of appropriate combinations of the information contained in lexical entries (rather than just via the headword). Thus there is substantial understanding of the issues of scanning the publishers' typesetting tapes, parsing the individual entries, and normalising their content; and a number of projects are currently concerned with developing techniques and tools for 'opportunistic', unconstrained, search and browsing through on-line MRDs. Prominent among these are the University of Waterloo project for computerising the New OED (Stubbs and Tompa, 1984; Tompa, 1986), the development of the notion of a 'dictionary server' at Xerox PARC (Kay, 1984), and the system for creating and querying multiple dictionaries under development at IBM Yorktown Heights (Neff *et al.*, 1988); in Europe similar concerns are at the focus of lexical projects based at Amsterdam (van der Steen, 1982), Cambridge (Boguraev *et al.*, 1987) and Pisa (Calzolari, 1988).

A characteristic property of the majority of work reported here is that it is carried out with a number of different (mostly English) MRDs which include *Webster's Seventh Collegiate Dictionary*, OED, LDOCE, OALD, Collins (mono- and bilingual), the Italian Zingarelli (Zingarelli, 1970) and Garzanti (Garzanti, 1984), the Collins Thesaurus (Collins, 1984), and others. Most groups, in fact, have focussed more or less exclusively on one MRD; notable exceptions here are the work by the group in Pisa, who are beginning to analyse the issues in setting up an integrated bilingual lexical database system, by linking a number of monolingual databases through a bilingual MRD (Calzolari and Picchi, 1986);

and the work that has been carried out by the Lexical System Group at IBM (Byrd *et al.*, 1987).

The fact that different MRD sources are used at different sites means that the lexical databases developed by individual research groups do not necessarily share properties, and almost certainly cannot be merged as a matter of course. An LDB is a highly specialised software system, designed to impose some structure on a machine-readable dictionary. The kind of a dictionary (and dictionary format) a particular group is working with, the balance between structured data and free text to be found in it, the particular information sought from it, and the nature of the NLP application program(s) which require this information impose stringent constraints on the design of the associated LDB. So far, very few efforts have been aimed explicitly at designing a 'dictionary interchange format' and standardising the functionality of, and software systems for creating, lexical databases (such concerns, however, are beginning to be voiced by a number of established lexical research groups in the United States — see, for instance, Amsler, 1987, and the recent workshop on *Text Encoding Standard for the Humantities* at Vassar College, Sperberg-McQueen, 1987).

This proposal from the perspective of current research The main motivation of this proposal is to integrate European research on MRDs. We intend to build on existing research by bringing together the major European groups actively working with MRDs for the purposes of NLP applications. Furthermore, we aim to pool together lexical, as well as human, resources. Such a synergy is clearly methodologically desirable, as it will bring together a number of research groups which are motivated by essentially similar interests, ultimately pursue the same goal, and have complementary skills and experience. The synergy is also practically necessitated by the fact that no single dictionary, particularly in a multilingual context, can be reasonably expected to provide all the information to be encoded in the lexical database and knowledge base we aim to construct.

The project will address a number of outstanding problems in the lexical capabilities of NLP systems; a lexical knowledge base constructed from a number of MRDs would go a long way towards resolving the problem of limited vocabulary and would provide an operational definition of general lexical and conceptual knowledge, which would in turn provide the basis for the semi-automatic acquisition of domain-specific vocabulary and knowledge for particular application systems. Furthermore, a multilingual MRD with a common conceptual structure would solve the same problems in the context of multilingual applications, such as translation, and would provide the basis for efficient and rapid transportability of monolingual NLP systems between languages as well as application domains. Finally, as Wilks *et al.* (1988) point out, dictionaries and knowledge representation schemes share common (essentially hierarchical) organising principles, therefore detailed research with MRDs may yield further insights into the problems and issues in the representation and manipulation of knowledge in general.

4 Outstanding problems with MRD work and our approach to them

In the course of our research we intend to address a number of outstanding problems of computational lexicography, some of which are alluded to in the previous section. To date, the bulk of research with MRDs has been carried out from sources of *English dictionaries*; furthermore, very little explicit effort has been expended on the utilisation of mono- and bilingual dictionaries in a *multilingual context*. A particular consequence of such decentralised work is not only the repeated (low-grade) activity of mounting the same source(s) at different sites, but the costly development of *ad-hoc database systems and access procedures*, typically non-applicable to other on-line resources, and non-generalisable to different access requirements. Another consequence is the *specificity of existing techniques* for extraction of lexical data: currently such techniques are closely tied to a particular language (English) and (individually) applicable to a small range of MRDs. Finally, a particular weakness of machine-readable sources of published dictionaries, from the point of view of developing data extraction techniques and building computational lexicons, is their *lack of complete reliability* along several dimensions.

4.1 Beyond a single language

In marked contrast to current work with MRDs, we will carry out parallel studies of a number of different monolingual dictionaries available in machine-readable form. The particular languages which the proposers will be working with are Dutch, English and Italian; lexical resources to explore include LDOCE, Collins (mono- and bilingual), Zingarelli, van Dale (mono- and bilingual), COBUILD and the Longman Roget's Thesaurus (1983).

Within a language, we intend to analyse the structures of individual lexicons and develop algorithmic procedures for extracting lexical data relevant to the automated processing of that particular language; we also expect to recover lexical regularities across languages. Commonalities in the definitional spaces of a cross-section of dictionaries will suggest a conceptual structure underlying general principles for knowledge representation; this structure will then be used as the organising principle of the lexical knowledge base.

Across languages, we intend to use multilingual dictionaries for at least two purposes. Firstly, while there are strong arguments in favour of a single conceptual structure underlying different languages, it is not at all clear that this is systematically reflected in bilingual dictionaries. Such structure would undoubtedly be of considerable utility to MT systems, particularly those within the paradigm of *knowledge-based* machine translation. We intend to carry out detailed studies, by bringing together bilingual dictionaries with their monolingual counterparts, of the notion of a shared knowledge base, constructed on the basis of core defining concepts across (mono-lingual) dictionaries, and of the feasibility of its extraction on the basis of existing (mono- and bilingual) MRDs. Typically, complementary pairs of bilingual dictionaries (e.g. English–Italian and Italian–English) display a non-trivial amount of asymmetry. In certain cases, such asymmetry is motivated by sound lexicographic principles, and one of the implications of this fact is that a multilingual LDB/LKB cannot be derived by a simple 'union' of (its) monolingual counterparts: thus the parallel analysis of bilingual sources, in order to bridge the (conceptual) gap between monolingual LDBs, will be an integral part of the process of multilingual LDB creation.

Secondly, we regard bilingual dictionaries as vehicles for transfer of lexical information between languages, where some data, extracted from a monolingual source and tied in to particular word-senses, needs to be associated with the corresponding items at (the lexical

level) in a lexicon for the target language. An example of such a procedure would be the mapping of selectional restrictions/semantic markers, extracted from an English source and carried over to a lexicon for an Italian parsing system. As we pointed out earlier, such transfer is necessary in order to make the kinds of information, for the extraction of which a particular mono-lingual dictionary is best suited, equally available not only in (other) monolingual LDBs, but across languages as well.

4.2 A methodology for utilising (unreliable) MRDs

The computational lexicography literature lists a number of ways in which printed dictionaries are (erroneously) incomplete: missing grammatical information, insufficient details concerning the context warranting the use of a particular word, incomplete (and, in particular, 'loose') definitions of word meanings, are only a few examples of such *errors of omission*. Likewise, not all instances of asymmetry across pairs of bilingual dictionaries are justified: omissions of words in one dictionary, where their translations are listed in its counterpart, constitute just another example of errors of omission on the part of the lexicographers.

A different way in which dictionary sources are unreliable manifests itself through *errors of commission*: examples here range from erroneous information assigned to (aspects of) a word sense to simply typographic errors which have slipped through the publishers' proof-reading procedures. A special case in this category is a common weakness, shared by virtually all dictionaries, due to the utilisation of various 'formal' systems for the presentation of certain lexical data. Even though such formats appear, or claim to be, consistent, often the conventions used by lexicographers to present such data are either not formal enough to allow easy decoding by machine, or abused (deliberately, for the sake of better visual presentation or as a space saving device, or accidentally, due to human error and lack of rigorous checks), or both. The result of such errors are syntactically incoherent and/or semantically inconsistent entries; circularity in dictionary definitions or very loosely structured set of core defining terms are particularly good examples here (see Boguraev and Briscoe, 1988b, for fuller examples).

Given the proven utility of MRDs as sources of lexical data, the question then arises how to make maximal use of them, notwithstanding their inherent unreliability. Our approach to this is based on the development of a particular methodology for utilising on-line dictionaries, which promotes a strong separation between the notions of *extracting* information from on-line sources and *using* this information for practical purposes.

In practical terms, this will be achieved by introducing a level of neutrality in the output of the extraction programs. Particular extraction procedures will be defined to produce 'proto-entries' in a form which can subsequently be piped into a whole family of related applications. The derivation of computational lexicons and/or LDBs from such an intermediate representation is incorporated within an environment implementing rapid and semi-automatic generation of lexical information in a suitable format, and monitoring the correctness/appropriateness of the resulting data (Boguraev *et al.*, 1987; and Carroll and Grover, 1988, discuss one particular system designed to behave in such a way).

Being able to posit theory-neutral intermediate representations is not only a desirable prerequisite for true sharing of lexical data on a large scale. In the context of the search for an appropriate methodology for using machine-readable sources, such an organisation of the transition between the processes of extraction and use of lexical data allows for softly configurable systems, in which fine tuning of both the extraction software and the shape of the back-end lexical knowledge base can proceed independently.

4.3 Towards a computational model of a dictionary

As research into analysing MRDs intensifies, rectifying the inedaquacies of existing systems for making MRDs available on-line becomes more urgent. In particular, not enough attention has been paid to date to developing systems which are general enough for loading and manipulating different dictionaries, existing in different machine-readable formats. Furthermore, access schemes have not only been weak in respect of expressive power of the queries they support, but also strongly dependent on the particular dictionary format and language. Finally, very little effort has gone into developing systems capable of simultaneously holding more than one dictionary on-line and allowing cross-MRD/cross-language comparisons and studies. (The only, notable, exception here is the work at IBM Yorktown Heights Research Centre, where researchers have developed a software package — *WordSmith* — giving on-line access to over 10 MRDs. However, this system is not fully general, because it adopts idiosyncratic representational schemes for the individual dictionaries: see, for example, Byrd and Neff, 1987).

Fundamentally, all these inadequacies are due to the same factor: lack of coordination of research activity, and consequently little consensus, on the issue of a suitable computational representation of a dictionary. There are good reasons for not using conventional database management systems for storing and accessing lexical databases; there are, however, very few detailed proposals concerning alternative access methods. A convergent view is that of representing lexical data by hierarchical structures which mirror the logical organisation of an entry in a source dictionary (Neff *et al.*, 1988; Calzolari, 1988). However, this representational scheme has only been applied to a limited number of MRDs to date, and more research is neede to ascertain whether this is the optimal computational model of a dictionary.

We propose to investigate these issues by experimenting with a range of data models derived from applying advanced computer science techniques to the particular constraints imposed by concerns of representation, storage, retrieval and manipulation of dictionary data. We will develop a general purpose system capable of parsing a typesetting tape according to an explicitly specified grammar of lexical entries. We will then design and implement a query language, with appropriate interfaces both for on-line (interactive) browsing and for off-line (batch) retrievals of large volumes of lexical data. We will augment the system with capabilities of coercing, or merging, data derived from individual MRDs into a common, shared, lexical database. Finally, we will investigate the appropriateness of the data model we evolve for representation of lexical data to the task of encoding, on a large scale and in a form tractable to NLP systems, the lexical knowledge in the LKB we aim to build.

The MRD sources we will be primarily working with are only one example of large volumes of text in electronic form; encyclopedias, thesauri, term banks, document collections and text corpora, for example, also fall in this category. Indeed, we already share experience in analysing large text corpora for the purposes of discovering regularities in language, particularly when they are of relevance to NLP systems (see, for example, the work in Lancaster on the Lancaster/Oslo-Bergen corpus: Briscoe *et al.*, 1986; Garside *et al.*, 1987); in deriving tagged corpora from machine-readable dictionaries (Akkerman and Meijs, 1987, 1988); and in improving access to large text corpora using the organisational structure of (MRD-derived) lexical databases (Calzolari and Picchi, 1988). The representational techniques and processing tools we develop will be sufficiently general to be applicable, in different contexts, to such textual databases; in turn, access to such

large corpora can provide vital statistical information on, for instance, the relative spread and frequency of different meanings of lexical units, their syntagmatic and domain-linked behaviour, and so forth.

4.4 Evaluation of lexical data extracted from MRDs

A large proportion of work with MRDs is still in a pre-theoretical, exploratory stage. With a few exceptions, where very specific data is sought from a particular dictionary to support a well-defined task (e.g. deriving a word list for a spelling corrector, or assigning syntactic features to words as required by a parser), activity has been centred on 'opportunistic' use of dictionaries, and the focus of such activities has been the feasibility of extracting, for example, selectional restrictions for verbs or the genus terms from word sense definitions. Indeed, these are precisely the kinds of investigation which underpin the current belief that machine-readable sources are of enormous utility to (applied) computational linguistics — noone would dispute the value of semantic markers or defining concepts in building natural language systems.

However, the real extent to which information extracted from an MRD can be utilised by such systems is still unclear. Essentially, there are two related issues. Firstly, since no single dictionary source can be reasonably expected to contain, and provide, all the lexical information that might be required in different contexts and applications, it is important to know how far the extraction procedures can be pushed in specific cases. Indeed, this is not only the motivation for building a common, shared lexical database by incrementally compiling and merging data acquired from separate sources; it is also necessary to gauge the limitations of such sources individually, in order to proceed with a consistent, and optimal, design for the lexical database.

Secondly, as we pointed out earlier, it is not clear that all of the conceptual information required by NLP system can be derived in a systematic and consistent way from dictionary sources. Some statement, then, concerning the functional completeness of a lexical database/lexical knowledge base from the perspective of applied computational linguistics, as opposed to purely theoretical linguistics and lexicography, is going to be of immense benefit to the NLP community.

We propose to investigate these questions by experimenting with particular applications designed to make use of our prototype lexical knowledge base. In addition to being 'vehicles' for evaluating the degree of usefulness of the LKB, we will regard the requirements, both at the lexical and knowledge level, of these applications as a guide on the content of a lexical resource of such scale and intent. We have chosen generic language processing tasks which require non-trivial 'understanding' of their input text(s), and thus will fully exercise the capabilities of the LKB. Furthermore, the development of the applications — natural language query interpretation and knowledge-based machine translation — will provide the context, both monolingual and multilingual, for addressing questions like depth and breadth of conceptual coverage of the lexical knowledge base, its real suitability for coping with the knowledge acquisition bottleneck, its commonality across different languages, and its precision in allowing/facilitating the mapping between surface words and underlying concepts.

5 Research objectives

The research we propose to carry out is driven by the very specific lexical needs of (knowledge-intensive) NLP programs. It will provide context for a coordinated study of a representative sample of MRDs, both within and across a number of European languages, from the perspective of their utility for supporting a range of mono- and multilingual language processing applications. The goal of constructing a multilingual LKB will require detailed studies of the linguistic similarities and differences in the lexical systems of different languages, as well as a reconsideration, from a very pragmatic perspective, of some central issues in knowledge representation. It will promote the development of a family of tools and techniques for loading, accessing and manipulating MRDs on-line, with particular emphasis on supporting the algorithmic extraction of different kinds of lexical data; the design of these facilities will be grounded in a special purpose data model, suitable for the representation of dictionary databases.

The overall, long-term, goal of our research is the development and construction of a lexical knowledge base, neutral with respect to individual languages, rooted in a common conceptual structure underlying a 'deep' processing model of language, and containing a substantial general vocabulary. The range of lexical data in the LKB — from phonological to semantic/pragmatic — should make it capable of supporting the (lexical) knowledge requirements of a variety of practical NLP systems.

Below we list some specific objectives of this work.[5]

Computational model of an MRD This will have to be defined as a sufficiently general representation capable of handling a variety of different dictionaries (e.g. different publishers'/typesetting formats), in different languages (e.g. English, Italian and Dutch), and with different aims (e.g. monolingual learner, monolingual reference, bilingual translation). Following the development of a special-purpose data model for an on-line dictionary, generic data storage and access functions will be implemented, in parallel with a system for parsing 'raw' dictionary sources and converting them into lexical databases with a common format.

Tools and techniques for converting an on-line MRD into an LDB These include an additional level of lexicographically specialised software, built on top of the programs discussed above. While a number of such tools will necessarily be tied to particular dictionary formats, the outcome of their application will be classes of formally represented lexical data, encoded in a suitable 'theory- and application-neutral' format. Appropriate formats will need to be designed for the various types of information to be encoded in the lexical entries of the LDB/LKB. These include, for example, phonological information, information about subcategorisation and valency, or the meaning structure(s) of dictionary definitions.

Tools and techniques for merging LDBs Such techniques will use constraints from linguistic theory, as well as a certain amount of extra-linguistic knowledge. In this context we will look at questions like how to achieve sense identification and mapping between

[5]Note that a number of deliverables are directly applicable to tasks and contexts which are peripherally related to ours, but independently equally central to concerns of computational lexicography (e.g. design of systems for creating new dictionaries) or natural language processing (e.g. development of general knowledge representation schemes).

dictionaries, what constitutes sufficient evidence in support of merging information from different sources, and how to resolve conflicts (due to e.g. errors of commission or omission).

A different dimension to the same objective is the development of tools and techniques for deriving a multilingual LDB from monolingual ones. We view this process as considerably more complex than a simple extension to the one above, since it crucially relies on a detailed study of linguistic regularities and differences between the languages involved. The target multilingual LDB is not a new, and separate, entity from the source monolingual ones. It is more in the nature of a 'super-structure', rich in inter-connections which define the correspondences between the monolingual sources. A bilingual dictionary in this context is not only the source from which these connections are derived, but also the media through which the inter-language transfer is achieved.

The task of merging individual LDBs is going to provide one particular framework for evaluating our notion of a 'theory-neutral' representation, from the perspective of its suitability for encoding data from different sources (dictionaries and languages).

Regularities in lexical systems of different languages The question of the degree of similarity between the lexical systems of European (Germanic and Romance) languages is both theoretically interesting and of great practical importance for the construction of a multilingual LKB. The LKB will aim to integrate the most useful and reliable information from a variety of MRDs. In a monolingual context, the transfer and integration of information is relatively straightforward; for example combining the subcategorisation codes of LDOCE with the COBUILD synonym and hyponym information could be done by establishing a mapping between word senses in the two dictionaries. It is plausible that this mapping could be derived on the basis of relatively crude definitions analysis programs, such as Alshawi's (Alshawi, 1988). However, in a multilingual context, mapping between senses (via a bilingual dictionary) does not guarantee that information can successfully be transferred to, for instance, Italian verbs with equivalent meanings, because there is no guarantee that, say, a verb taking a sentential (*that*) complement in English (*say, understand, think, believe*) will take one in Italian (see Byrd, Calzolari *et al.*, 1987). On the other hand, since such syntactic properties of verbs are closely connected to their meaning we would expect there to be substantial overlap.

Furthermore, transferring such information would be of considerable practical benefit, because no Italian MRD contains the kind of detail concerning subcategarisation information offered by LDOCE. In this case, transfer would need to be guided by knowledge concerning semantic classes of verbs with similar syntactic behaviour (see Levin, 1988). In general, many such issues will need investigation using the MRDs contained in the multilingual LDB and its associated query language to obtain the relevant data.

Evolution of an LDB into a Lexical Knowledge Base This is closely tied in with techniques for the identification and representation of lexical relations between words (or, more precisely, word-senses): for the purposes of NLP systems, these relations must include IS-A, PART-WHOLE, synonymy, USED-FOR, and so forth. This is the general conceptual lexical knowledge, which, while implicitly present in virtually all dictionaries, is one of their most elusive aspects because it is hidden by the sequential alphabetic organisation of their printed versions (Calzolari, 1984a, 1988).

The issue here is not so much that of developing new techniques, but improving existing ones: for instance, if a prototype definitions analysis system, based on simple pattern matching techniques, can achieve 80% success in parsing dictionary entries, we can clearly

expect substantial improvement utilising current state-of-the-art parsers.

Another issue concerns the minimal functional properties of such a lexical knowledge base. These include, in particular, its 'neutral' aspect, which specifically stipulates its being *multi-functional* and *multi-perspective*. It is clearly desirable for an LKB to be made accessible, via its reorganisation into a fully functional database with a front end incorporating a mechanism akin to Kay's dictionary server (Kay, 1984), to a number of different applications. It is also necessary to be able to obtain different 'snap-shots', or 'cross-sections' of it, as an application requires, for instance, information about the subcategorisation properties of a verb sense, followed by a traversal of the IS-A hierarchy, followed by the extraction of a synonym set. Such questions have already been subject to preliminary study (Calzolari, 1988; Byrd, Calzolari *et al.*, 1987; Byrd, 1988), in a monolingual context, but much work remains to be done.

General conceptual lexical knowledge The work under this heading is going to provide the theoretical foundation for the development of a (multilingual) LKB, since it addresses one of the fundamental questions, already raised several times in this document and underlying, implicitly, the proposed research: how closely, and how adequately, does the semantic structure of a dictionary (and its definitions) 'slot into' currently accepted knowledge representation frameworks — particularly the ones adopted by practical NLP systems. In particular, this part of the research is going to be responsible for establishing the exact set of semantic/conceptual relationships (IS-A, PART-WHOLE, USED-FOR and so forth, above) underlying the LKB structure. Related issues will be covered by a search of a mapping of word-sense definitions across different dictionaries, and a study of the feasibility of replacing (or enhancing) definitions in one dictionary, e.g. Zingarelli, with those of another, e.g. LDOCE.

In this context we intend to develop techniques for identifying core sets of (language specific) defining concepts, and eliciting the particular knowledge structures they populate. Next, we shall seek commonalities between these 'proto-lexicons', in order to elicit a shared taxonomy of defining concepts used in the process of dictionary writing and to confirm, empirically, the other basic claim, to be explored and tested in this research, that such a taxonomy is common across different languages.

Strong arguments in favour of organising the definitions of the core set of concepts into a hierarchical semantic structure inlude the facts that such a structure not only reflects the natural divisions of lexical conceptual organisation, but also facilitates more economic definitions employing property inheritance (see e.g. Flickinger *et al.*, 1985). Furthermore, in addition to the convenience it offers for analytic term definition, which is the basis for defining more complex concepts as composites of simpler ones, such a taxonomy serves a purpose beyond simply providing a versatile way of organising a knowledge base. Suitably enhanced with an automatic classification procedure, it also allows the maintenance of a *dynamically changing* knowledge base — a critical requirement for any knowledge structure which has to support robust language processing in realistic environments, where novel or unfamiliar user inputs are not exceptional.

Case studies: the utility of an LKB An evaluation of our approach to extracting lexical knowledge from MRDs and making this available to NLP applications will be carried out by experimenting with two largely extant prototype systems, built on the basis of existing natural language processing software (Cater, 1986, 1988; Grover *et al.*, 1988; Phillips and Thompson, 1987; Russell *et al.*, 1986; Crabtree, Crouch, Pulman *et al.*, 1988).

The choice of language processing tasks below reflects both our perception of practical contexts to which advances in computational linguistics are of particular relevance, and our concern that all different aspects of the LKB are put to the test.

We intend therefore to evaluate the utility of the LKB in a multilingual, as well as a monolingual, environment: it will be used as an integral part of the design and function of a *machine translation* system and a *query interpretation* system.

In addition to the specific issues discussed below, the process of evaluating the LKB will provide a complementary perspective on the evaluation of the notion of 'neutrality' of form and content of the lexical knowledge, namely its suitability to, and accessibility by, different applications performing different tasks. We aim to develop a knowledge base which will be equally appropriate to frameworks using widely differing theories of semantic processing, and it is in this sense that the neutral semantic representation ought to be as easy to transform into, for instance, a semantic network incorporating inheritance and defaults, as into a set of meaning postulates (expressed in the notation of, say, First Order Logic).

Both prototypes will be targeted at specific domains, thus testing the suitability of the LKB for providing and building up the domain-specific knowledge typically required for realistic text processing. A knowledge base of the kind we propose would be capable of supporting rapid construction of a domain lexicon — for instance, a core set of domain terms supplied to the system could easily be integrated into the knowledge base (if not already there) and/or expanded by selectively following synonym/hyperonym links. Furthermore, both prototypes will be instances of systems which fall within the 'deep understanding' paradigm of computational linguistics, thus putting to test the suitability of the LKB for supporting general reasoning and inference procedures relying on access to structured knowledge about the world.

In addition, the individual tasks will further test other aspects of the LKB. Thus the requirements of machine translation provide a framework for evaluating its suitability to lexicon acquisition in a 'multi-language' context. In particular, we will be looking at the capability of the knowledge base to support incremental, top-down, semantically guided derivation of transfer dictionaries. The assumption is that by virtue of the way in which it is derived — namely by controlled merging of monolingual LDBs, via an organising structure of core defining concepts and using a bilingual MRD as a bridging device — a conceptually organised knowledge base will facilitate the mapping between source and target lexical terms. Furthermore, due to the same factors, the 'on-line' availability of such a lexical knowledge base to an MT system during its operation will further facilitate processes like lexical selection and coherent discourse generation.

Likewise, the very nature of natural language front end systems typically requires them to cope with certain amount of unfamiliar words; thus our second prototype application will not only test the LKB in the task of domain vocabulary acquisition *prior to* such systems being put in operation, but also in the task of 'assimilating' new words, *after* a system goes 'on-line'. The underlying hypothesis here is that, via their definitions within the LKB (and by a process usually referred to as *classification*), domain concepts appropriate to the (domain) interpretation of novel lexical items can be naturally incorporated into the front end system's domain-dependent knowledge structures. This would achieve immediate robustness of performance, whose side effect would be an incremental enrichment of the knowledge base.

Both the above prototype applications are designed to test the LKB within the context of text processing. However, we also hope to test (particularly the phonological and

morphological/syntactic components of) the LKB in a speech-based query system too. This system is the subject of another ESPRIT proposal[6], and is intended to develop a speech driven interface to office automation software. We expect that the products of this research will consistently enhance the lexical capabilities of the latter system which is an ESPRIT deliverable scheduled for 1995 (i.e. 2 years after the end of this project).

[6]Proposal for an ESPRIT2A Project II.4.3, submitted to the Commission in April 1988: *Continuous Speech Understanding System for Extendable Vocabulary*.

6 Programme of work; action management

Below is a chart of work plan and work load distribution between participants. Naturally, the lines between individual headings are not as clear-cut as they may appear; likewise, individual groups' participation is not necessarily strictly confined as presented. The chart mostly reflects experience and primary interests.

	Ams'm	C'bdg	D'bln	Lancs	Pisa
Computational model of an MRD	*	*			*
General conceptual lexical knowledge	*		*	*	*
Design / desiderata for LKB		*		*	*
Migration of LDB into LKB		*	*		*
Methodologies / tools / techniques:					
- MRD => LDB	*	*		*	*
- merging LDBs	*			*	*
- LDB => LKB		*		*	*
- LKB => applications		*	*		
Regularities in lexical systems across different languages	*			*	*
Lexical requirements of					
- multilingual NLP systems			*		*
- monolingual NLP systems				*	*
Case studies: utility of LKB					
- multilingual context			*		
- monolingual context		*			

The programme is staged over three years, with work on individual projects proceeding more or less in parallel at the different sites. The ordering of the headings in the chart above reflects the overall (even though not exact) sequence in which these will be undertaken.

More specifically, the first year will be devoted to experimenting with different computational models for on-line dictionaries and devising a common format, from which a suitable internal representation of individual entries will be developed. During this period, and following the finalisation of the standard dictionary format, we will also implement the necessary software for converting an MRD into an LDB.

At the beginning of the second year, and after the MRD-to-LDB conversion system has been distributed across the sites, we will commence work on merging the individual LDBs. We expect that the processes of merging monolingual and multiligual LDBs will have to be phased due to the time it is going to take to, for instance, carry out the studies required for establishing how certain lexical properties (e.g. subcategorisation, semantic classes of verbs, selectional restrictions) map across languages. Since such studies are necessarily dependent on having a monolingual LDB, itself the result of merging several dictionary databases, it will probably take of the order of six months to acquire the information necessary for the initial multilingual LDB 'bootstrap'.

The same studies will be also used for the analysis of semantic structure of dictionaries: thus during the second half of the second year we will also begin the investigation into the general conceptual knowledge, underlying the definitions in dictionaries and to be encoded into the lexical knowledge base. We expect by the end of the second year to have

identified the core semantic/conceptual relationships, and thus be ready for the migration of the LDB into an LKB.

Shortly within the third year, after an initial prototype LKB is available, we will begin its evaluation. This is clearly going to be an iterative process, as we will have to incrementally adapt and/or enhance the LKB format and content, as we get a better understanding of the lexical requirements of NLP systems in both monolingual and multilingual contexts. While some of this latter work can be carried out independently, and in parallel with the evolution of the LKB from the original dictionary sources, it can only be completed by observing the needs for lexical knowledge of the two prototypes (MT/query interpretation). There is also going to be some overlap between the mainstream work in computational lexicography and LKB construction and the development of the prototype test application systems — as pointed out earlier, these will not be built completely from scratch, but will largely utilise already existing tools and components.

Similarly, work on studying regularities in lexical systems across languages can proceed, to a certain degree, independently and in parallel with the initial development of the MRD-to-LDB software. All of the groups concerned already have access to on-line MRDs, and the only prerequisite for being able to begin these studies is an agreement on the type of linguistic phenomena to focus on. The availability of multilingual lexical databases thus is only a requirement for the latter part of this particular project, when generalisations will have to be drawn about the nature of the mappings of these phenomena across languages.

Internal assessment of progress will be carried out as part of preparing the 6-monthly reports required by the Commission; given the programme phases, as outlined above, review points fall naturally at year boundaries. In order to achieve, and maintain, coordination between the individual groups, the coordinating site (University of Pisa/Institute for Computational Linguistics), will hold regular progress meetings; furthermore, the Pisa budget (see Appendix A), incorporates financial support for part-time managerial staff.

7 Deliverables

The results of the proposed research programme can be summarised in terms of the following three broad categories:

Written materials Above and beyond (internal departmental) working papers and research reports, as well as the six-monthly progress reports required in the context of Basic Research Actions, we will produce a number of publications in a variety of forms. The most common way of making our work known to the community will be via research papers, published in conference proceedings (to achieve swift dissemination of results) and refereed journals (to achieve widest circulation). In addition, we will jointly produce and publish books in the broader area of computational lexicography. These may arise as edited collections of papers presented at specialist workshops (see below), in the style of Walker, Zampolli and Calzolari (1988); alternatively they may be specially written to report on the particular background and progress of our research, as in Boguraev and Briscoe (1988a). In this way we intend to fill a large gap in the currently available literature in the field (indeed, the two volumes cited above are currently the only ones addressing the concerns of computational lexicography from the perspective of computational linguistics/natural language processing). Finally, we expect that opportunities will arise to present lectures to the academic community: as a result, we will be producing a number of publications of a survey and/or expository nature (thus a number of proposers are already involved in teaching at a forthcoming summer school on *Computational Lexicography and Lexicology*, organised by the European Science Foundation and hosted by the Institute of Computational Linguistics in Pisa).

Direct dissemination of knowledge/training aspects Presenting lectures is just one way of making the results of our research available. We will also organise annual workshops primarily for project personnel but also for other researchers directly involved in this field and potential 'clients' of our research, such as NLP system designers and commercial users. These workshops will assist with the dissemination of experience within the project and the further training of research staff. To this end, the application includes a request for funding of travel and subsistence for project members between the five sites. This funding will also support some individual visits of extended duration between sites to facilitate the integration and coordination of the research effort. Finally, we expect researchers in the member institutions not directly involved in the project to receive indirect benefit and training through the medium of Ph.D theses and so forth, undertaken in the general area of our proposed research programme.

Specifications of prototype software systems In addition to standard dissemination of research results via academic publication, we expect to produce and make available to the research community specifications of:

- a standard dictionary interchange format,
- a general-purpose computational representation of an MRD,
- a set of core tools and techniques for converting, accessing and manipulating on-line MRDs and lexical databases, and
- the form and content of a lexical knowledge base.

The software systems developed during the course of the project will be available to all the consortium members for their continuing research.

Finally, we will present a methodology for utilising potentially rich, but inherently unreliable, MRD sources; and will offer an evaluation of the notion of 'theory-neutral', as it applies to the concerns of applied computational lexicography and natural language processing.

8 References

Akkerman E (1988) A independent analysis of the LDOCE grammar coding system. In Boguraev B, Briscoe E (eds) *Computational Lexicography for Natural Language Processing*. Longman, Harlow and London, pp 65–84.

Akkerman E, Meijs W J, Voogt-van Zutphen H J (1987) Grammatical tagging in ASCOT. In Meijs W J (ed) *Corpus Linguistics and Beyond, Proceedings of the Seventh International ICAME Conference*. Rodopi, Amsterdam, pp 181–93.

Akkerman E, Meijs W J, Voogt-van Zutphen H J (1988a, forthcoming) *A Computerized Lexicon for Word Level Tagging: ASCOT Report No 2*. Rodopi, Amsterdam.

Alshawi H A (1988) Analysing the dictionary definitions. In Boguraev B, Briscoe E (eds) *Computational Lexicography for Natural Language Processing*. Longman, Harlow and London, pp 153–170.

Amsler R A (1980) 'The structure of the Merriam-Webster Pocket Dictionary'. Doctoral thesis, University of Texas at Austin.

Amsler R A (1981) A taxonomy for English nouns and verbs. *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics,* Stanford, California, pp 133–8.

Amsler R (1987) 'An interchange format for machine-readable dictionaries', Record of a Working Group on Knowledge and Data Bases, Workshop on the Lexicon in Computational and Theoretical Perspective, Stanford CA

Atwell E S (1981) 'LOB Corpus tagging project: manual pre-edit handbook'. Departments of Computer Studies and Linguistics, University of Lancaster (unpublished).

Atwell E S (1982) 'LOB Corpus tagging project: manual postedit handbook (a minigrammar of LOB Corpus English, examining the types of error commonly made during automatic (computational) analysis of ordinary written English)'. Departments of Computer Studies and Linguistics, University of Lancaster (unpublished).

Atwell E S (1983) Constituent-likelihood grammar. *ICAME News* 7: 34–67.

Atwell E S, Leech, G.N. and Garside, R.G. (1984) Analysis of the LOB corpus: progress and prospects. In J Aarts and W Meijs (eds) *Corpus Linguistics*. Rodopi, Amsterdam, pp 41–52.

Atwell E S, Elliott S (1987) Dealing with ill-formed English text. In Garside R, Leech G, Sampson G (eds) *The Computational Analysis of English: a Corpus-Based Approach*. Longman, London and New York, pp 120–138.

Beale A D (1985a) A probabilistic approach to grammatical analysis of written English. *Proceedings of the Proceedings of the 2nd Conference of the European Chapter of the Association for Computational Linguistics,* Geneva, pp 159–65.

Beale A D (1985b) Grammatical Analysis by Computer of the Lancaster-Oslo/Bergen Corpus. *Proceedings of the Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics,* Chicago, pp 293–298.

Beale A D (1988) Lexicon and Grammar in Probabalistic Tagging of Written English. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Buffalo, New York.

Boguraev B K (1987) The definitional power of words. *Proceedings of the 3rd Workshop on Theoretical Issues in Natural Language Processing (TINLAP-3),* Las Cruces, New Mexico, pp 11–15.

Boguraev B K (1988, forthcoming) Machine readable dictionaries and research in computational linguistics. In Walker D, Zampolli A, Calzolari N (eds) *Automating the*

Lexicon: Research and Practice in a Multilingual Environment. Cambridge University Press, Cambridge.

Boguraev B K, Briscoe E J (1987) Large lexicons for natural language processing: exploiting the grammar coding system of LDOCE. *Computational Linguistics* **13**(3–4).

Boguraev B, Briscoe E (eds) (1988a) *Computational Lexicography for Natural Language Processing*. Longman, Harlow and London.

Boguraev B, Briscoe E (1988b) Computational Lexicography: Introduction. In Boguraev B, Briscoe E (eds) *Computational Lexicography for Natural Language Processing*. Longman, Harlow and London, pp 1–40.

Boguraev B, Briscoe E (1988c) Utilising the LDOCE grammar codes. In Boguraev B, Briscoe E (eds) *Computational Lexicography for Natural Language Processing*. Longman, Harlow and London, pp 85–116.

Boguraev B K, Briscoe E J, Carroll J, Carter D, Grover C (1987) The derivation of a grammatically indexed lexicon from the Longman Dictionary of Contemporary English. *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, Stanford, California, pp 193–200.

Boguraev B K, Carter D M, Briscoe E J (1987b) A multi-purpose interface to an online dictionary. *Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics*, Copenhagen, Denmark, pp 63–9.

Booth, B.M. (1985) Revising CLAWS. *ICAME News* **9**: 29–35.

Briscoe E J (1985) *Report of the Dictionary Syndicate*. Alvey Speech Club Workshop, Warwick University.

Briscoe E J, Craig I, Grover C E (1986) 'The use of the LOB corpus in the development of a phrase-structure grammar of English'. In *Proceedings of the Seventh International Conference on English Language Research on Computational Corpora (ICAME)*, Amsterdam, The Netherlands.

Byrd R J (1988, forthcoming) Dictionary systems for office practice. In Walker D, Zampolli A, Calzolari N (eds) *Automating the Lexicon: Research and Practice in a Multilingual Environment*. Cambridge University Press, Cambridge.

Byrd R, Calzolari N, Chodorow M, Klavans J, Neff M, Rizk O (1987) 'Tools and methods for computational lexicology'. RC-12642, IBM Yorktown Heights.

Calzolari N (1984a) Detecting patterns in a lexical database. *Proceedings of the Tenth International Congress on Computational Linguistics*, Stanford, California, pp 170–3.

Calzolari N (1984b) Machine-readable dictionaries, lexical databases and the lexical system (panel session on Machine-Readable Dictionaries). *Proceedings of the 10th International Conference on Computational Linguistics*, Stanford, California, pp 460.

Calzolari N (1988, forthcoming) Structure and access in an automated lexicon and related issues. In Walker D, Zampolli A, Calzolari N (eds) *Automating the Lexicon: Research and Practice in a Multilingual Environment*. Cambridge University Press, Cambridge.

Calzolari N, Antona M (1987) 'Proposal for a classification of Italian verbs in a lexical database'. Working Paper, Laboratorio di Linguistica Computazionale, CNR, Pisa.

Calzolari N, Picchi E (1986) 'A project for a bilingual lexical database system'. In *Proceedings of the Second Annual Conference of the University of Waterloo Centre for the New OED*, Waterloo, Canada.

Calzolari N, Picchi E (1988) 'Textual perspectives through an automatized lexicon'. Working Paper, Laboratorio di Linguistica Computazionale, CNR, Pisa.

Carroll J A, Grover C E (1988) The derivation of a large computational lexicon of English from LDOCE. In Boguraev B, Briscoe E (eds) *Computational Lexicography for Natural*

Language Processing. Longman, Harlow and London, pp 117–134.

Carter D M, Boguraev B K, Briscoe E J (1987) Lexical stress and phonetic information: which segments are most informative. *Proceedings of the European Conference on Speech Technology*, Edinburgh, pp 235–8.

Carter D M (1988) LDOCE and speech recognition. In Boguraev B, Briscoe E (eds) *Computational Lexicography for Natural Language Processing*. Longman, Harlow and London, pp 135–152.

Cater (1986) Preference-directed use of ATNs. *Proceedings of the European Conference on Artificial Intelligence (ECAI-86)*, Brigton, pp 23–29.

Cater (1988) 'A robust semantic analyser for English'. Working paper, Department of Computer Science, University College Dublin, Dublin.

Chodorow M S, Byrd R J, Heidorn G E (1985) Extracting semantic hierarchies from a large on-line dictionary. *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago, Illinois, pp 299–304.

Chodorow M S, Ravin Y, Sachar H E (1988) A tool for investigating the synonymy relation in a sense disambiguated thesaurus. *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, TX, pp 144–151.

Church K (1985) Stress assignment in letter-to-sound rules for speech synthesis. *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago, Illinois, pp 246–54.

Collins (1984) *The New Collins Thesaurus*. Collins Publishers, Glasgow.

Collins (1980) *Collins Sansoni Italian Dictionary: Italian–English/English–Italian*. Collins Publishers, Glasgow.

Crabtree B, Crouch R S, Moffat D C, Pirie N, Pulman S G, Tate A (1988) A natural language interface to an intelligent planning system, paper to be presented at the *National Information Technology (Alvey) Conference*, University College Swansea, UK.

Flickinger D, Pollard C, Wasow T (1985) Structure-sharing in lexical representation. *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago, Illinois, pp 262–267.

Fox E A, Nutter J T, Ahlswede T, Evens M, Markowitz J (1988) Building a large thesaurus for information retrieval. *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, TX, pp 101–108.

Garside R, Leech F (1985) A probabilistic parser. *Proceedings of the Proceedings of the 2nd Conference of the European Chapter of the Association for Computational Linguistics*, Geneva, pp 166–170.

Garside R, Leech G, Sampson G (eds) (1987) *The Computational Analysis of English: a Corpus-Based Approach*. Longman, London and New York.

Garzanti (1984) *Il Nuovo Dizionario Italiano Garzanti*. Garzanti, Milano.

Gazdar G, Klein E, Pullum G, Sag I (1985) *Generalized Phrase Structure Grammar*. Basil Blackwell Publisher Ltd, Oxford.

Grover C, Briscoe E, Carroll J, Boguraev B (1988) 'The Alvey natural language tools project grammar — a large computational grammar of English'. Lancaster papers in Linguistics (No. 47), Department of Linguistics, University of Lancaster.

Haas N, Hendrix G (1983) Learning by being told: acquiring knowledge for information management. In Michalski R, Carbonell J, Mitchell T (eds) *Machine Learning: an Artificial Intelligence Approach*. Tioga Publishing Company, Palo Alto, California, pp 405–26.

Hornby A S (ed) (1980) *Oxford Advanced Learner's Dictionary of Current English* 3rd edition (11th impression). Oxford University Press, Oxford.

Huttenlocher D P (1985) *Exploiting sequential phonetic constraints in recognizing spoken words*. Massachussetts Institute of Technology Artificial Intelligence Laboratory AI. Memo 867.

Huttenlocher D P, Zue V W (1983) Phonotactic and lexical constraints in speech recognition. *Proceedings of the National Conference on Artificial Intelligence (AAAI-83)*, Washington, DC, pp 172–6.

Ingria R (1984) 'Complement types in English'. Report No. 5684, Bolt Beranek and Newman Inc, Cambridge, Massachusetts.

Johansson S, Leech G, Goodluck H (1978) 'Manual of Information to accompany the Lancaster/Oslo-Bergen Corpus of British English for use with digital computers'. Department of English, University of Oslo.

Johansson S, Atwell E, Garside R, Leech G (1986) 'The Tagged LOB Corpus Users' Manual'. Norwegian Computing Centre for the Humanities, Bergen.

Kay M (1984) The dictionary server (panel on Machine-Readable Dictionaries). *Proceedings of the 10th International Conference on Computational Linguistics (Coling84)*, Stanford, California, pp 461.

Knuth D E (1986) *The TEXbook*. Addison Wesley Publishing Company, Reading, MA.

Leech G, Garside R, Atwell E S (1983a) The automatic grammatical tagging of the LOB Corpus. *ICAME News* 7: 13–33.

Leech G, Garside R, Atwell E S (1983b) Recent developments in the use of computer corpora in English language research. *Transactions of the Philological Society*, pp 23–40 1983b

Lenat D, Prakash M, Shepherd M (1986) CYC: using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine* 6(4): 65–92.

Levin B (1988, forthcoming) Approaches to lexical semantic representation. In Walker D, Zampolli A, Calzolari N (eds) *Automating the Lexicon: Research and Practice in a Multilingual Environment*. Cambridge University Press, Cambridge.

Liang F M (1983) 'Word Hy-phen-a-tion by Com-put-er'. PhD thesis, Report STAN-CS-83-977, Stanford University Computer Science, Stanford, CA.

Markowitz J, Ahlswede T, Evans M (1986) Semantically significant patterns in dictionary definitions. *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, New York, pp 112–19.

Marshall, I (1983) Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB Corpus. *Computers and the Humanities* 17: 139–50.

Michiels A (1982) 'Exploiting a large dictionary database'. Doctoral thesis, Université de Liège, Belgium.

Meijs W, den Broeder M, Vossen P (1988) Meaning and structure in dictionary definitions. In Boguraev B, Briscoe E (eds) *Computational Lexicography for Natural Language Processing*. Longman, Harlow and London, pp 171–192.

Neff M S, Byrd R J (1987) 'WordSmith users guide'. IBM Research Report, T J Watson Research Center, Yorktown Heights, New York.

Neff M S, Byrd R J, Rizk O A (1988) Creating and querying lexical data bases. *Proceedings of the Second ACL Conference on Applied Natural Language Processing*, Austin, Texas, pp 84–92.

Phillips, John and Thompson, Henry (1987) A parser and an appropriate computational representation for GPSG, in Klein E and Haddock N (eds), *Cognitive Science Working*

Papers 1, Centre for Cognitive Science, University of Edinburgh

Procter P (ed) (1978) *Longman Dictionary of Contemporary English.* Longman Group Ltd, Harlow.

Roget's Thesaurus (1973) *Roget's Thesaurus of English Words and Phrases.* Longman.

Russell G, Pulman S, Ritchie G, Black A (1986) A dictionary and morphological analyser for English. *Proceedings of the 11th International Congress on Computational Linguistics*, Bonn, Germany, pp 277–279.

Sinclair J, Editor-in-Chief (1987) *The Collins COBUILD English Language Dictionary.* William Collins Sons & Co. Ltd., London.

Sperberg-McQueen M (1987) 'Text encoding standard for the humantites — Vassar workshop report', to appear in ACM SIGIR Forum

Shipman D, Zue V (1982) Properties of large lexicons: implications for advanced isolated word recognition systems. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, pp 546–9.

van der Steen G J (1982) A treatment of queries in large text corpora. In Johansson S (ed) *Computer Corpora in English Language Research.* Norwegian Computing Centre for the Humanities, Bergen, pp 49–65.

Stubbs J, Tompa F (1984) *Waterloo and the New Oxford English Dictionary Project.* Paper presented to the Twentieth Annual Conference on Editorial Problems, University of Toronto, Toronto.

Tompa F (1986) *Database design for a dictionary of the future.* Preliminary report, Centre for the New Oxford English Dictionary, University of Waterloo, Waterloo, Ontario.

van Dale (1984) *Van Dale Groot Woordenboek.* Van Dale Lexicografie, Utrecht/Antwerpen.

Voogt-van Zutphen H J (1987) 'Constructing an F.G. lexicon on the basis of LDOCE'. (Working Papers in Functional Grammar No.24), Dept. of Linguistics, University of Amsterdam.

Voogt-van Zutphen H J (1988, forthcoming, forthcoming) Towards a Lexicon of Functional Grammar. In Connolly J and Dik S (eds) *Functional Grammar and the Computer.* Foris, Dordrecht.

Vossen P (1988, forthcoming, forthcoming) The meaning descriptions in the lexicon provided by the LINKS project. In Connolly J and Dik S (eds) *Functional Grammar and the Computer.* Foris, Dordrecht.

W7 (1967) *Webster's Seventh New Collegiate Dictionary.* C.&C. Merriam Company, Springfield, Massachussetts.

Wahrig G (1986) *DTV Wörterbuch der Deutschen Sprache.* Deutscher Taschenbuch Verlag, München.

Walker D, Amsler R (1986) The use of machine-readable dictionaries in sublanguage analysis. In Grishman R, Kittredge R (eds) *Analyzing Language in Restricted Domains.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp 69–83.

Walker D, Zampolli A, Calzolari N (eds) (1987) 'Towards a polytheoretical lexical database'. Preprints, Laboratorio di Linguistica Computazionale, CNR, Pisa.

Walker D, Zampolli A, Calzolari N (eds) (1988) *Automating the Lexicon: Research and Practice in a Multilingual Environment.* Cambridge University Press, Cambridge.

Whitelock P, Wood M, Somers H, Johnson R, Bennett P (eds) (1987) *Linguistic Theory and Computer Applications.* Academic Press, New York.

Wilks Y, Fass D, Guo G, McDonald J, Plate T, Slator B (1988) A tractable machine dictionary as a resource for computational semantics. In Boguraev B, Briscoe E (eds)

Computational Lexicography for Natural Language Processing. Longman, Harlow and London, pp 193–228.

Yannakoudakis E, Fawthrop D (1983) The rules of spelling errors. *Information Processing and Management* 19(2): 87–99.

Zampolli A (Pisa, Italy) Lexicological and lexicographical activities at the Instituto di Linguistica Computazionale. In Zampolli A, Capelli A (eds) *The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries*. Report on an International Workshop on "Computers in Dictionary Production", pp 237–278. 1984

Zampolli A, Picchi E, Calzolari N (1987) The use of computers in lexicography and lexicology. In Cowie A (ed) *The Dictionary and the Language Learner*. Lexicographica Series Maior 17, Niemayer, Tubingen, pp 55–77.

Zingarelli N (1970) *Vocabolario della Lingua Italiana*. Zanicelli, Bologna.

A Administrative/financial details; site budgets

Common equipment requirements

A characteristic property of this project is the necessary sharing, on a large scale, of code which will be incrementally and modularly developed at different sites. An additional requirement is being able to exchange, on a regular basis, large volumes of textual data and knowledge structures: for example, the 'raw' source of an MRD can easily reach 20–30 MBytes; after formatting and converting it to a database, it can double up in size; a lexical knowledge base, derived ultimately by merging individual LDBs, is estimated to be even larger.

A number of the partners are situated in Linguistics departments, and thus have somewhat limited access to computing equipment; in particular, while they may be able to arrange access to the University mainframes, and thus both use existing prototype systems and carry out the bulk of raw text processing work in batch mode, they will not be able to use the advanced 'lexicographer's workstation' software, as described in section 6 (*Deliverables*), and developed during the lifetime of the project.

Therefore, in order to ensure ultimate software compatibility between sites, as well as to promote portability of raw lexical resources and formatted databases, the individual proposed expenditures presented below budget for low-cost personal workstations (whose operational characteristics include: sufficient amount of internal RAM; virtual memory support; large, detachable and transportable hard discs, TCP communications boards, and capable of running a European-flavoured Common Lisp).

TABLE II-1 BR/CFP 88

1. PROPOSAL SUMMARY INFORMATION

Action Title: Acquisition of Lexical Knowledge
for Natural Language Processing Systems

Action Area: Artificial Intelligence and Cognitive Science / Computer Science

Action Duration: 36 months

2. TOTAL EXPENDITURE BREAKDOWN

	Proposer	Expenditure (ECUs)
Coordinator	A.Zampolli/N.Calzolari	525 000
Partner 2	B.Boguraev	350 000
Partner 3	E.Briscoe	347 309
Partner 4	A.Cater	262 585
Partner 5	W.Meijs	262 500
	Total expenditure	ECUs 1 747 394