Lexical Issues in Natural Language Processing

Ted Briscoe

University of Cambridge Computer Laboratory
New Museums Site, Pembroke Street, Cambridge, CB2 3QG, UK

November 1991

(in E. Klein and F. Veltman (eds.), *Natural Language and Speech,* Springer-Verlag, 1991)

ESPRIT BRA-3030 ACQUILEX Working Paper No 041

Lexical Issues in Natural Language Processing

Ted Briscoe
Computer Laboratory, University of Cambridge
Pembroke Street, Cambridge, CB2 3QG, UK
ejb@cl.cam.ac.uk

1 Introduction

In this paper, I will briefly describe the role of the lexicon in natural language processing (NLP) applications and will go on to discuss a number of issues in lexical research and in the design and construction of lexicons for practical NLP applications. I will survey relevant research in Europe, America and Japan; however, in a paper of this length it is not possible to consider every instance of a particular approach, so neither the text nor references should be taken to be exhaustive.

In recent years, the lexicon has become the focus of considerable research in (computational) linguistic theory and NLP; the reasons for this trend are both theoretical and practical. Within linguistics, the role of the lexicon has increased as more and more linguistic generalisations have been seen to have a lexical dimension. Within NLP, the lexicon has increasingly become the chief 'bottleneck' in the production of habitable NLP systems offering an adequate vocabulary for the intended application. This has led to the use of machine-readable versions of conventional dictionaries in an attempt to develop substantial lexicons for NLP in a resource efficient fashion[1]. At the same time, dictionary publishers and lexicographers have realised the potential benefits of information technology in the dictionary production process and also the potential new markets for their products that might be created by commercial development of language technologies. This latter possibility creates the very exciting opportunity for collaborative research and development between these apparently rather disparate communities which, I believe, would substantially remove the existing bottleneck for NLP systems and would provide a major impetus to theoretical research on the lexicon.

In the next section, I review relevant research in recent linguistic theory and describe the highly-structured, hierarchical and 'generative' conception of the lexicon which is emerging from this work. I then go on to discuss the advantages and disadvantages of exploiting machine-readable dictionaries and progress which has been made, to date, on constructing substantial lexicons with such resources. I also discuss several new projects which, in reaction to this trend, have opted for manual construction. In the fourth section, I provide more detailed examples illustrating some of these trends and developments, drawn from the ACQUILEX project, through which I and my colleagues have been both exploring the utility of machine-readable dictionaries for NLP and contributing to the development of a theoretically-motivated, but substantial and computationally-tractable, multilingual lexicon. In the conclusion, I discuss the more active role that lexicographers and dictionary publishers have begun to play in lexical research and suggest how this development might be harnessed to facilitate solutions to outstanding problems in lexicon design and development.

2 The Lexicon in Linguistic Theory

At least since Bloomfield (1933), the lexicon has usually been viewed as the repository of idiosyncratic and unpredictable facts about lexical items organised as an (alphabetic) list; for example, that *kill* in English means 'x cause y to die' and is a transitive verb with regular morphology. On the other hand, the fact that the subject of *kill* appears before it in typical English sentences or that its past participle is *killed* were taken to be predictable and quite general statements about English syntax and morphophonology which should be stated independently of the lexicon. However, within generative linguistic theory since the 1970s there has been a consistent increase in the

[1] I shall use the term 'dictionary' to refer to the conventional printed object for human use and 'lexicon' for a formal and possibly implemented dictionary intended either as a component of linguistic theory or of an NLP system.

role of the lexicon in capturing linguistic generalisations, both in the sense that more and more of the rules of grammar are coming to be seen as formal devices which manipulate (aspects of) lexical entries and in the sense that many of these rules are lexically-governed and must, therefore, be restricted to more finely specified classes of lexical items than can be obtained from traditional part-of-speech classifications. As the importance of the lexicon has increased, so the role of other components of the overall theory of grammar has declined; thus in some contemporary theories, the syntactic component is reduced to one or two general principles for the combination of constituents, whilst all the information concerning the categorial identity and mode of combination of these constituents is projected from individual lexical entries. However, this shift of emphasis makes it increasingly difficult to view the lexicon as a simple list of lexical entries (like a conventional dictionary) since this organisation does not support generalisation about classes of lexical items.

2.1 Lexical Grammar

Chomsky (1970) discussed the problem of nominalisation for generative (transformational) grammar and proposed a new theory of grammar in which lexical redundancy rules rather than transformations were used to express the relationship between a verb and a morphologically derived nominal. Chomsky's arguments for this move were, in part, concerned with the restriction of transformational operations, but also with the idiosyncratic properties of many derived nominals; that is, the rules which relate derived nominals to their corresponding verbs are often semi-productive because, for example, although the morphological operation involved in the derivation of the nominal is regular, its meaning is specialised and unpredictable (*revolve, revolution*). Work following this influential paper has tended to emphasise the semi-productive or lexically-governed nature of many other phenomena and to make greater use of formal devices, such as lexical redundancy rules, which serve to relate lexical entries and enrich the structure of the lexicon. The introduction to Moortgat et al. (1980) provides a detailed account of these developments.

One landmark in the development of lexical grammar is the account of subcategorisation and coordination developed within Generalized Phrase Structure Grammar (GPSG, Gazdar *et al.*, 1985). Simplifying somewhat, lexical items are subcategorised via a feature which effectively indexes them to specific Phrase Structure (PS) rules which introduce their appropriate syntactic arguments as phrasal sisters; for example, the PS rules in (1a,b) are appropriate expansions for transitive verbs which take a noun phrase (NP) object and for verbs taking infinitival verb phrase (VP) objects, respectively.

(1) a VP $\rightarrow$ V[Subcat 2] NP[Acc]
b VP $\rightarrow$ V[Subcat 6] VP[Infin]
c X $\rightarrow$ X X[Conj and]

Verbs of each type will be listed in the lexicon with appropriate values for the Subcat feature, so *kill* would have the value 2, whilst *try* would be 6. GPSG also posited very general PS rule schemata for coordination of which a simplified example is given in (1c), for binary conjunction where X ranges over syntactic categories. These rules interact together to predict the (un)grammaticality of the examples in (2) in a simple and intuitive way.

(2) a Kim [$_{VP}$ [$_{VP}$ killed Sandy] [$_{VP}$ and tried to leave]]
b Kim killed [$_{NP}$ [$_{NP}$ Sandy] [$_{NP}$ and her friend]]
c Kim tried [$_{VP}$ [$_{VP}$ to pay] [$_{VP}$ and to leave]]
d *Kim killed [$_{?}$ [$_{NP}$ Sandy] [$_{VP}$ and to leave]]

Thus, coordination is constrained by lexical projections in the form of PS rules indexed to lexical (sub)classes via the Subcat feature. However, the search for a unified account of local grammatical agreement lead Pollard (1984) to propose a framework in which the Subcat feature takes as value an ordered list of syntactic categories and PS rules are replaced by a very general PS schema which combines a lexical item with the topmost category of its Subcat list and creates a new phrasal category with a 'popped' Subcat list. In this framework, the (simplified) lexical entry for a transitive verb would be (3a) and the PS schema would construct the analysis outlined in (3b) (where I abbreviate the Subcat feature to its value).

(3) a kill : V[Subcat <NP[Acc] NP[Nom]>]
b $[_{V[<>]}$ Kim $[_{V[<NP[Nom]>]}$ killed him]]

Agreement of features such as person, number or case (illustrated here) can be straightforwardly and uniformly enforced by encoding such features on categories specified on the Subcat list of lexical items. Within this framework, the syntactic component has been drastically reduced since individual PS rules have been replaced by a single schema which builds constituents according to the specifications of Subcat lists projected from lexical entries. This schema will, however, interact with that for coordination in (1c) to cover the examples illustrated in (2.)

One apparent disadvantage of this radically lexical approach to grammar is that it appears to involve considerable redundancy and loss of generalisation if the lexicon is organised as an unrelated list of entries; for example, (3a) encodes the information that the subject of *kill* combines with it after its object and that the subject must be nominative. However, these facts generalise to all verbs of English whilst the fact the *kill* takes only one object generalises to all transitive verbs. Further developments of syntactic theory have reinforced the trend to relocate information in the lexicon (e.g. Pollard & Sag, 1987; Steedman, 1985; Zeevat *et al.*, 1987). Flickinger *et al.* (1985), Pollard & Sag (1987) and others propose that the lexicon be represented as an inheritance hierarchy in which information common to a class of lexical items is inherited by all its subclasses; thus, the information that verbs take nominative subjects is associated with the verb class node and inherited by all subclasses, such as transitive verbs. These proposals enrich the struture of the lexicon in a fashion which allows generalisations about lexical (sub)classes to be expressed economically.

2.2 Lexical Semantics

Most formal semantic theories have concentrated on the problems of compositional rather than lexical semantics; that is, the construction of sentence meaning from the meaning of constituent words and phrases. Many lexical theories of grammar are monostratal, admitting only one level of syntactic representation (e.g. Gazdar *et al.*, 1985). These theories associate semantic representations with each syntactic constituent in some fashion; for instance, early versions of GPSG paired a semantic rule with each syntactic PS rule, which built the semantics of the left-hand mother category out of the semantics of each right-hand daughter category. Within the more radically lexical theories, the compositional semantics of (at least) lexical items and their syntactic arguments is also relocated in the lexicon. Such theories are often called sign-based because they formally instantiate Saussure's concept of a linguistic sign as the (arbitrary) association of sound, form and meaning (e.g. Pollard & Sag, 1987). In a sign-based theory, the lexical entry for a transitive verb will include the information that the semantics of the subject and object syntactic arguments function as the semantic arguments of the predicate associated with the verb. This information too generalises to all transitive verbs, but locating it in the lexicon allows the same very general schema which is used to construct the syntactic representation of phrases and clauses to also build up the semantic representation in tandem. (An example of a lexical entry of this type is given in the next section.)

Recently, there has been renewed interest and research on the meaning of words themselves and, in particular, work on how lexical semantic properties affect both syntactic behaviour and compositional semantic interpretation. To take just two examples: Levin (1988, 1990) has argued that it is not adequate to simply list alternative syntactic realisations of verbs in terms of separate lexical entries with distinct values for the Subcat feature or its equivalent, because such alternate realisations are partly predictable on a semantic basis and may have semantic consequences. For instance, change of possession verbs such as *give* often undergo the dative alternation illustrated in (4a,b).

(4) a Kim gave the beer to Sandy
b Kim gave Sandy the beer
c Kim slid the beer to Sandy / the table edge
d Kim slid Sandy / *the table edge a beer

Change of position verbs such as *slide*, however, can only undergo the dative alternation if they can be interpreted as conveying a change of possession, as (4d) illustrates. Pustejovsky (1989a,b) discusses examples such as (5a,b,c) in which *enjoy* conveys an identical relationship of pleasurable experience between the experiencer subject and an event denoted by the verb's object of which the experiencer is agent.

(5) a Kim enjoys making films
b Kim enjoys film-making
c Kim / Coppola enjoyed that film

Positing separate lexical entries on the basis of the differential syntactic realisations of *enjoy* with either a NP or progressive VP object fails to capture the semantic relatedness of these examples; thus, in (5b) we need to account for the manner in which the implicit agent of the event-denoting NP *film-making* is associated with Kim, whilst in (5c) we must explain the mechanism which allows *that film* to denote an event of Kim watching (or Coppola making) a film. Pustejovsky refers to this latter process as logical metonymy since he argues that *enjoy* coerces its artifact-denoting NP object into an event of some type, whilst the lexical semantic representation of the NP itself determines the broad nature of the understood event — compare *Kim enjoyed a beer*.

Work of this type on lexical semantics, as well as much other research on, for example, aksionsart (Sanfilippo, 1990), or the argument structure of derived nominals (Grimshaw, 1990), poses a considerable challenge to lexical grammar and theories of lexical organisation. Such theories must demonstrate how lexical semantic information affects patterns of syntactic realisation and also the process of compositional interpretation.

2.3 The Lexical Representation Language

Theories of grammar must be expressed in a formal language with an appropriate syntax and semantics. As theories have become more lexical, the focus of such metatheoretical work has also shifted to the lexicon. I will refer to the language in which the lexicon is expressed as the lexical representation language (LRL). Most monostratal and lexical theories of grammar treat syntactic categories as feature structures (FSs) with unification as the mode of combination of information in FSs. Unification is a form of bi-directional pattern matching which is used extensively in theorem proving and logic programming and which owes its introduction into linguistic theory as much to work in NLP (Kay, 1974; Shieber, 1984) as to work on theories of lexical grammar. A FS for the transitive verb *kill* is given in (6) which could constitute (the syntactic and semantic part of) its lexical entry in a theory of the type outlined in the last two sections. This FS is displayed in attribute-value matrix notation in which coindexing indicates token identity of subparts of the FS and bold face expressions give the type of each (sub)-FS (see 4.4 below). Unification of two FSs, if defined, produces a new FS in which the information from both is monotonically combined. Shieber (1986) and Kasper & Rounds (1990) provide detailed introductions to unification-based approaches to grammar and to the syntax and semantics of the formalism.

The FS in (6) is simple by comparison to that which would be required in a realistic, wide-coverage grammar, yet already it encodes a large amount of information much of which is true of other transitive verbs. The LRL should allow the aspects of this FS common to all transitive verbs to be expressed just once rather than repeated in each individual lexical entry. Shieber (1984) describes the use of lexical templates to name and define subparts of FSs common to classes of lexical items and to abbreviate entries themselves to lists of template names which would be expanded to full FSs on demand. This approach compacts lexical entries and allows the expression of certain generalisations, particularly as templates can be embedded within other template definitions. However, templates are abbreviatory devices which do not enforce any specific organisation of the lexicon and which do not strongly constrain the featural content of FSs. Moens *et al.* (1989) present a typed FS system and Carpenter (1990, 1991) develops a scheme in which FSs are typed and structured via a partial order on types. Thus, a type places appropriateness conditions on a class of FSs and these conditions must also be satisfied by FSs which are subtypes of that type. The type system can be used to define an inheritance hierarchy in which FSs 'lower' in the type hierarchy are monotonically enriched with information derived from appropriateness conditions.

$$
(6)\quad \left[\begin{array}{l}
\textbf{strict-trans-sign}\\
\mathrm{ORTH} = \textbf{kill}\\
\mathrm{CAT} = \left[\begin{array}{l}
\mathrm{RESULT} = \left[\begin{array}{l}\textbf{strict-intrans-cat}\\ \mathrm{RESULT} = \textbf{sent-cat}\\ \mathrm{ACTIVE} = \left[\begin{array}{l}\textbf{np-sign}\\ \mathrm{SEM} = \boxed{1}\end{array}\right]\end{array}\right]\\
\mathrm{ACTIVE} = \left[\begin{array}{l}\textbf{np-sign}\\ \mathrm{SEM} = \boxed{2}\end{array}\right]
\end{array}\right]\\
\mathrm{SEM} = \left[\begin{array}{l}
\textbf{strict-trans-sem}\\
\mathrm{IND} = \boxed{3}\,\textbf{eve}\\
\mathrm{PRED} = \textbf{and}\\
\mathrm{ARG1} = \left[\begin{array}{l}\textbf{verb-formula}\\ \mathrm{IND} = \boxed{3}\\ \mathrm{PRED} = \textbf{kill}\\ \mathrm{ARG1} = \boxed{3}\end{array}\right]\\
\mathrm{ARG2} = \left[\begin{array}{l}
\textbf{binary-formula}\\
\mathrm{IND} = \boxed{3}\\
\mathrm{PRED} = \textbf{and}\\
\mathrm{ARG1} = \boxed{1}\left[\begin{array}{l}\textbf{p-agt-formula}\\ \mathrm{IND} = \boxed{3}\\ \mathrm{PRED} = \textbf{p-agt}\\ \mathrm{ARG1} = \boxed{3}\\ \mathrm{ARG2} = \textbf{obj}\end{array}\right]\\
\mathrm{ARG2} = \boxed{2}\left[\begin{array}{l}\textbf{p-pat-formula}\\ \mathrm{IND} = \boxed{3}\\ \mathrm{PRED} = \textbf{p-pat}\\ \mathrm{ARG1} = \boxed{3}\\ \mathrm{ARG2} = \textbf{animate}\end{array}\right]
\end{array}\right]
\end{array}\right]
\end{array}\right]
$$

It is possible to define several notions of lexical rule within this framework. Shieber (1984) describes a very general mechanism which can define arbitrary mappings between two FSs, whilst Copestake & Briscoe (1991) propose to constrain this somewhat by placing further conditions on 'input' and 'output' FS expressed in terms of a type system. Bresnan & Kanerva (1989) and Bresnan & Moshi (1989) make use of a more restrictive notion of lexical rule which only allows monotonic enrichment of underspecified lexical entries.

Many researchers have proposed to augment the logic of FSs with further operations such as negation, disjunction, conditional implication and default inheritance (e.g. Pollard & Sag, 1987; Carpenter, 1990, 1991; Russell *et al.*, 1990; Evans & Gazdar, 1990). Unconstrained addition of such operations considerably complicates either the computational tractability or the semantic interpretation of the formalism, or both, and there is still debate over which such extensions are linguistically and formally desirable (see e.g. the papers in Briscoe, Copestake & de Paiva, 1991). In a wider context, it is not clear that a restriction to unification-based formalisms is either tenable or desirable; Pustejovsky (1989a), for example, makes use of mechanisms drawn from general purpose knowledge representation languages developed within artificial intelligence to characterise certain lexical semantic processes. Hobbs *et al.* (1987) argue that capturing the semantics of lexical items requires a first-order logical representation and mechanisms which support abductive as well as deductive inference. Undoubtedly the inferential processes involved in language comprehension extend beyond the limited mechanisms provided within unification-based formalisms, however, it is not clear yet whether lexical operations *per se* require them.

3 The Lexicon and Natural Language Processing

Natural language processing applications, ranging from superficial text critiquing through to machine translation, require knowledge about words. In most cases, to be practical and habitable such systems must be furnished with a substantial lexicon covering a realistic vocabulary and providing the kinds of linguistic knowledge required for the application. Even apparently simple applications often require quite diverse lexical knowledge; for instance, it is straightforward to make a case for a spelling checker to be able to utilise orthographic knowledge, phonological knowledge (to deal with confusions caused by homonymy), morphological knowledge (if only to allow access to the lexicon), syntactic knowledge (to allow recognition of errors creating legal words in syntactically illegal contexts), crude semantic knowledge (to select specialised term banks), and so forth. In a survey of systems covering a range of tasks, Whitelock *et al.* (1987) found that the average lexicon size was 1500 lexical items (which fell to 25 if one large MT system was discounted). The Core Language Engine (Alshawi *et al.*, in press), a state-of-the-art unification-based parsing and generation system for English intended as a generic front-end for computer systems, has a lexicon of about 2000 words (Carter, 1989).

One reason why the lexical capabilities of NLP systems has remained weak is because of the labour intensive nature of encoding lexical entries. If we assume that the task of developing an adequate 'core' lexicon is equivalent to that of developing a conventional advanced learners' dictionary from scratch (containing typically 50,000 entries), then the labour required runs into hundreds of person / years. Furthermore, the *Oxford English Dictionary* contains 500,000 entries (and is still unlikely to cover all the words which will be found in naturally-occurring input (Walker & Amsler, 1986). As the sophistication and coverage of other aspects of NLP systems increases, so the need to address this problem becomes more urgent. Since the resources required for manual development of lexicons are typically not available, some NLP researchers have turned to the machine-readable versions of conventional dictionaries (MRDs) as a highly-structured and substantial source of lexical information. However, there are disadvantages to this approach because MRDs present information in a manner which relies on the linguistic skills and background knowledge of the user, whilst implementable theories of the lexicon can make no such assumptions. Therefore, following the lead of lexicographers themselves (e.g. Sinclair, 1987; Summers, 1990), others have opted to attempt to (semi-)automatically acquire lexical information from naturally-occurring textual and transcribed spoken corpora, whilst in several recent projects the emphasis has been placed squarely on manual development of lexicons by large teams of researchers.

3.1 Exploiting Machine-Readable Dictionaries

There are a wide variety of MRDs which are made available by dictionary publishers in formats ranging from typesetting tapes through to quite sophisticated databases. The ideal resource from the perspective of NLP research would be a fully explicit advanced learners' dictionary organised as a database which has undergone very systematic error checking. These sources tend to assume less linguistic competence on the part of the user and, therefore, supply more grammatical information, use restricted defining vocabularies, and so forth, but MRDs such as the *Longman Dictionary of Contemporary English* (LDOCE), the *Oxford Advanced Learners' Dictionary* (OALD) or *Collins Cobuild English Language Dictionary* only approach this ideal and as advanced learners' dictionaries are rare for languages other than English, considerable energy and some debate has been devoted to the problems of converting MRDs into databases, recognising and compensating for errors or inadequacies, and so forth. Early work tended to focus on a single MRD (e.g. Amsler, 1981), whilst recent efforts have attempted to merge and integrate information from more than one source and sometimes more than one language (e.g. Byrd *et al.*, 1987). There are now several alternative and well-developed approaches to deriving a lexical database (LDB) from a typesetting tape (e.g. Boguraev *et al.*, 1991) offering the ability to access classes of lexical items and entries on the basis of any of the information contained in one or more converted sources.

The extraction of substantial quantities of information from a LDB is in one sense trivial because once a MRD has been converted it is easy, for example, to list all the verb senses which contain the word *cause* in their definition. However, the usefulness of this activity can only be evaluated relative to theoretical proposals concerning how such information might be utilised and represented in the lexicon. More abstractly, the contents of the LDB retain whatever implicit semantics was intended by the lexicographers, whilst the utilisation of this information requires a demonstration that it can be related to a formal theory of the relevant domain and represented in an appropriate LRL. For example, there is no guarantee that two MRDs will use a label like 'vt' (transitive verb) in the same manner, so a demonstration of the utility of this information requires that we relate it to a theory of syntax and subcategorisation of verbs, represent the information in the LRL provided by this theory and provide evidence that correct predictions are made about the lexical items involved. For instance, the theory concerned might predict via a regular lexical operation that transitive verbs can undergo passive, but the implicit definition used by the lexicographers might well not have used this as a criterion of transitivity. In this case, a direct mapping of 'vt' into the LRL will result in examples such as *Kim was resembled* or *5 pounds was weighed by the book* being parsed or generated. This kind of problem is more common than might be expected, precisely because of the static character of a printed dictionary or LDB, as compared to the dynamic nature of a LRL which incorporates a theory of valid lexical operations. I refer to the end result of this process of correctly mapping information from a LDB into the LRL as a lexical knowledge base (LKB). A LKB is, in effect, an instantiated LRL, the ultimate goal of research in this area.

To date, the most successful work on the construction of LKBs from MRD sources has been based on utilising more codified information, such as headword orthography, part-of-speech codes, grammatical codes and pronunciation fields; for example, Church (1985) discusses the use of pronunciation fields in the construction of an LKB for text-to-speech synthesis and Boguraev & Briscoe (1989) contains several papers which evaluate and describe the use of LDOCE grammar codes in a LKB with subcategorised verbs, adjectives and nouns. Work on lexical semantics has for the most part resulted in more codified and accessible LDB representations rather than in genuine LKBs; in part this is because of the lack of a theoretical consensus on most aspects of lexical semantics, but it also reflects the greater difficulty of extracting useful information from dictionary definitions intended for human consumption. Various pattern matching and parsing tools have been constructed for recognising genus terms in definitions and the syntactico-semantic relationship between genus and differentiae (e.g. Amsler, 1980; Alshawi, 1989; Vossen, 1990a). Using such tools the information from dictionary definitions has been structured in various ways to create, for instance, hierarchically structured taxonomies of genus terms, but often the senses of the genus terms have remained unresolved and where they have been resolved this has been in terms of the source dictionary's sense distinctions (e.g. Copestake, 1990; Guthrie *et al.*, 1990).

3.2 Manual Development of Large Lexicons

Several researchers have argued strongly that MRDs are inappropriate sources of information for a LKB because they are too far removed from any adequate theory of the lexicon (e.g. Gross, 1984). A larger number would maintain that the overhead of converting MRD sources into LDBs is too great given the often unreliable and unsystematic nature of the information that can be derived from them. Experience in the past with manual creation of large lexicons is difficult to evaluate, for instance the Linguistic String Project (Sager, 1981) is said to have developed a lexicon of about 10,000 word forms manually but no analysis of the accuracy of the entries or resources required to develop it is available. Recently, several projects (EDR, Japan, Uchida, 1990; GENELEX, Esprit, Normier & Nossin, 1990; MULTILEX, Esprit, McNaught, 1990) have begun in which the intention appears to be to develop quite substantial lexicons or, at least, LDBs primarily manually.

The EDR project will cost 100 million US dollars, run for 9 years and intends to develop bilingual resources for English and Japanese containing 200,000 words, term banks for 100,000 words, and 400,000 concepts defined in terms of a semantic network. Although development will be assisted by corpus analysis and by software support, the primary method of creating entries will be manual encoding by teams of researchers. The EDR project will undoubtedly advance the state-of-the-art in the production of 'electronic' dictionaries or in my terms LDB, but it is unclear to me whether the project will produce a LKB. This LDB will be of use to researchers in NLP, but I doubt that it will form a satisfactory base for direct deployment in most applications. In the descriptions of the project available to me it appears that the emphasis is entirely on achieving substantial coverage and not on the many theoretical issues which need addressing before it will be possible to develop a genuine LKB. In addition this project, as with the development of a conventional dictionary, will be prey to problems of inconsistency, errors of commission, and so forth, created by the use of teams of manual encoders. In this respect, it is surprising that there appears to be no link with a dictionary publisher and no attempt to exploit the considerable experience of lexicographers in the management of such projects (Atkins, 1990).

3.3 The Role of Corpora

Summers (1990) reminds us that Dr. Johnson's great dictionary of English published in 1755 was based on quotations from literature, rather than introspection about word meaning. The availability of vast machine-readable corpora (e.g. Liberman, 1991) and of software tools for analysing and selecting subsets of such corpora makes the task of empirical lexicography considerably easier. The Cobuild dictionary was developed by a team of lexicographers using a written corpus of 6 million words supported by software tools such as concordancers, a database and editors (Clear, 1987). Cobuild is, in many respects, superior to its predecessors, such as in recognising senses of words that had slipped through the introspective net of lexicographical tradition. Nevertheless, the Cobuild project was very resource intensive, because the task of analysing large quantities of unsorted citations of particular word forms in context is complex and time consuming (e.g. Fillmore & Atkins, 1991).

Recently, some NLP researchers (e.g. Brent, 1991; Hindle & Rooth, 1991) have advocated the use of corpora for automatic acquisition of lexical information. This raises the possibility of, at least, semi-automatic construction of lexicons directly from corpora. However, many fundamental problems in NLP will need to be solved before this highly desirable prospect becomes practical, because the extraction of many types of information from corpora usually presupposes the capability to automatically analyse the raw text in various ways. Furthermore, achieving this capability will itself involve developing substantial lexicons. For example, it would be useful to acquire information about the relationship between alternative senses of a word and its syntactic realisation, but how would one recognise an alternative sense or the syntactic and semantic relationships between it and the words and phrases in the surrounding context? This appears to require at least a theory of sense distinctions and a parser capable of phrasal analysis and of discriminating arguments from adjuncts which, in turn, implies the existence of a lexicon with reliable information about subcategorisation.

Nevertheless, robust techniques exist for some types of corpus analysis, such as part-of-speech tagging (e.g. de Rose, 1988) or derivation of surface collocations (e.g. Church & Hanks, 1990), and as work on statistical and robust approaches to corpus analysis continues more complex analysis will become reliable enough for routine use; for example, phrasal parsing (e.g. Hindle & Rooth, 1991). And already these techniques allow the derivation of information which in some respects surpasses that available from MRD sources; for example, information about the frequency of words occurring as different parts-of-speech. It seems both likely and desirable that corpus analysis will play a greater role in the acquisition of lexical information, but unlikely that this approach will supplant others or render more theoretical work irrelevant.

4 ACQUILEX

In this section, I will describe some of the research undertaken as part of the ACQUILEX project in order to make some of the earlier discussion more concrete. The goal of ACQUILEX is to demonstrate that information can be usefully extracted from multiple MRD sources in a resource efficient fashion in order to (eventually) construct a multilingual LKB for use in NLP applications. Work on ACQUILEX can be divided into two broad areas: firstly, the development of software tools and a database framework to facilitate the mapping from LDB to MRD, and secondly, the adoption or development of theoretical accounts of aspects of the lexicon and the subsequent construction of illustrative LKB fragments semi-automatically from LDBs using a further set of software tools design to integrate, transform and enrich LDB information.

4.1 Mapping from MRD to LDB

4.1.1 Functionality of the LDB

The LDB system developed at Cambridge implements the two-level dictionary model (Boguraev *et al.*, 1991). In the two-level model, the source dictionary is the primary repository of lexical data, and, separately from the dictionary source, sets of interrelated indices encode all statements about the structure and content of the data held in the dictionary. Thus all the information associated with the dictionary source is preserved, but structural relationships are also expressed. Since new indices can be added incrementally it is unnecessary to try and establish all the possible relationships from the start. This set-up is appropriate for a highly-structured but primarily textual object which is continuously having further structure imposed on it as more information is extracted / made explicit. The LDB can support the model of the Common Lexical Entry described in Calzolari *et al.* (1990) both in terms of the graphical presentation and underlying representation of MRD entries. The LDB system is used throughout the ACQUILEX project and by a number of other research groups, and is described in detail in Boguraev *et al.* (1989) and Carroll (1990).

The mounting of a new machine-readable dictionary in the LDB can be divided into four stages:

1. Transforming the MRD source into a suitable format, while preserving information. The complexity of this stage depends on the format of the tape supplied by the publisher.

2. Defining what the indices are and how they are to be extracted from entries.

3. Defining the format of queries that the user can construct (that is, the possible attributes and their hierarchical organisation), and how these queries correspond to the indices created for the dictionary.

4. Telling the system to create permanent files on disc holding the indices, and the menus for the graphical query interface.

In fact, two types of indices are created: one type on the contents of headword fields (and also optionally on internal entry sequencing information on the typesetting tape), enabling access to entries via their headwords (similar to the traditional way of using printed dictionaries); the other type based on the contents of entries, allowing the dictionary to be queried, and entries to be retrieved from it, on the basis of elements and their relationships within entries, rather than just by headword.

An LDB query consists of a hierarchical collection of attributes with associated values; for example the query

```
[[syn
   [gcode T1]]
 [sem
   [word show]]]
```

has two attributes at the top level: 'syn' and 'sem'; the attribute 'gcode' is beneath 'syn' with value 'T1', and 'word' beneath 'sem' with value 'show'. When looking up a query, the LDB, by default, computes the answers in a sense-based (rather than an entry-based) fashion; that is, it returns just the senses which satisfy the query, not the whole entry (unless of course all the senses in the entry satisfy it).

We also use the LDB to store information which is derived from the MRD source, but is not sufficiently analysed to make it part of a LKB. For example, the results of analysing the definitions, as described in the next section, are stored in the LDB as a *derived dictionary*, with entries which are in direct correspondence to the source dictionary. The LDB allows the user to apply a single query to two or more such corresponding dictionaries simultaneously.

4.1.2 Analysing definitions

It is not currently possible to produce an analysis of dictionary definitions using a conventional parser, with a general purpose grammar. Two approaches taken within the ACQUILEX project, have proved reasonably successful; using a robust, pattern-matching / parsing tool, and development of a special purpose grammar for definitions using a general purpose parser.

The flexible pattern matching / parsing tool (FPar) which is integrated with the Cambridge LDB is based on the system described in Alshawi(1989). This uses a grammar in the form of a hierarchy of patterns; the most general patterns providing some interpretation of a text even if the more specific and detailed ones fail. As an example of an (atypically detailed) use of FPar on LDOCE:

launch[3] a large usu. motor-driven boat used for carrying people on rivers, lakes, harbours, etc.

```
((CLASS BOAT) (PROPERTIES (LARGE))
  (PURPOSE
    (PREDICATION (CLASS CARRY) (OBJECT PEOPLE))))
```

FPar has also been applied to the Spanish VOX dictionary (Rodriguez *et al.*, 1990), which is a much larger dictionary than LDOCE, and does not make use of a restricted defining vocabulary. Rather than attempting to build grammars which would work for the whole of VOX, different pattern hierarchies have been developed for different semantically-related groups of definitions.

In contrast, general purpose parsers have been used with special purpose grammars. Vossen (1990) describes work on LDOCE which has now been extended to the Dutch Van Dale dictionaries; in Pisa, the IBM PLNLP system has been used (Montemagni, forthcoming). In these approaches it is necessary to develop quite complex specialised grammars both to deal with the sublanguage of definitions and to reduce the number of alternative analyses produced.

All these approaches allow the identification of genus phrases in definitions with a good degree of accuracy (better than 95% for Vossen's parser on LDOCE noun definitions). However some errors seem inevitable; for example in the LDOCE definition;

armadillo "a small animal native to the warm parts of the Americas"

the genus was identified as *native* by both Vossen and Alshawi's analysers. Analysis of the differentia is considerably more difficult and currently all these systems give only partial information concerning the syntactic and semantic relations which obtain between the phrases of the differentia and the genus term. Since such systems are time consuming to develop and tend to be specific to individual MRD sources, in the longer term utilising probabilistic and robust parsing techniques developed for corpus analysis seems desirable (see Briscoe & Carroll (1991) for a preliminary experiment with LDOCE definitions).

Although it is possible to produce a parsed structure which gives the genus phrase and some indication of the relationships in the definitions, this is not a disambiguated meaning representation, which is what is needed to reason about the content of the definitions to formally derive a lexical semantic representation in the LKB. Mapping from parsed definitions of this type to an LKB representation involves a combination of heuristics and user interaction, and is only possible within a context provided by the LRL. Copestake(1990) describes a program for producing taxonomies which can be used to provide an inheritance structure for the LKB. This program traces 'chains' of genus terms through the dictionary; thus starting from *animal* we might find *dog* and from *dog terrier*, and so forth. The essential steps in going from the definitions with the genus term identified to a structure which can be interpreted in the formal system are disambiguation of the genus term, and identification of the type of relationship which holds between the genus term and the definiendum. Consider the following LDOCE definition:

dictionary 1 a book that gives a list of words in alphabetical order, with their pronunciations and meanings.

Here *book* is the genus term, but it is essential to determine the sense used; other definitions use *book* in other senses, for example:

Genesis the first book of the Bible ...

User-specified heuristics are utilised to select the appropriate sense, such as degree of word overlap between the definitions of the senses of the genus and the current definition (see Copestake (1990) for further details). Sense disambiguation is done semi-automatically with the user confirming decisions concerning non-leaf nodes in the hierarchy which emerges. Typically, it takes about 1 hour to create an inheritance hierarchy for 500 word senses using LDOCE.

4.1.3 Correlating MRDs

Correlating MRDs is one way to overcome inadequacies, inconsistencies, omissions and the occasional errors which are commonly found in a single source (Atkins, 1991). Hopefully, integrating information from several sources will provide the missing information or allow errors to be detected.

To take an example, semantically-defined verb classes are instrumental in providing an indication of lexically-governed grammatical processes, such as the alternate syntactic realisations of the type discussed in 2.2, and should thus be included within a lexicon which supplied adequate information about verbs. For example, a verb such as *delight* should be specified as a member of the class of verbs which express emotion, i.e. psychological verbs. As is well known (e.g. Levin 1990), these verbs can be further classified according to the following parameters:

- affect is positive (*admire, delight*), neutral (*experience, interest*) or negative (*fear, scare*)
- stimulus argument is realized as object and experiencer as subject, e.g. *admire, experience, fear*
- stimulus argument is realized as subject and experiencer as object, e.g. *delight, interest, scare*

Unfortunately, conventional dictionaries do not supply this kind of information with consistency and exhaustiveness, so the technique of creating derived dictionaries where the information contained in the MRD is made more explicit is unhelpful in this case. For example, one approach would be to derive a dictionary where verbs are organized into a taxonomy by genus term, as in 4.1.2. Unfortunately, the genus of verb definitions is usually not specific enough to supply a taxonomic characterization which would allow the reliable identification of semantic verb classes. In LDOCE, for example, the genus of over 20% of verb senses (about 3,500 verb senses) is one of 8 verbs *cause, make, be, give, put, take, move, have*; many of the word senses which have the same genus belong to distinct semantic verb classes. This is not to say that verb taxonomies are of no value; nevertheless, the achievement of adequate results requires techniques which reclassify entries in the same source MRD(s) rather than making explicit the classification 'implicit' in the lexicographer's choice of genus term. Such a reclassification can be carried out by augmenting a conventional learner's dictionary with thesaurus information; thesauri provide an alternative semantically-motivated classification of lexical items, and are, therefore, natural candidates for the task at hand.

In the general case, the integration of information from distinct MRD sources is probably going to remain an unsolved problem for quite some time. This is simply because dictionaries seldom describe the same word using the same sense distinctions. Consequently, the integration of information from distinct MRD sources through simple word-sense matches is likely to fail in a significant number of instances (e.g. Calzolari & Picchi 1986). Indeed, Atkins & Levin (1990) have suggested that the task of mapping MRDs onto each other is so complex that the creation of a complete 'ideal' database, which provides a reference point for the MRD sources to be integrated, may well be an essential prerequisite. However, when dealing with MRD sources which use entry definitions which are not too dissimilar, a correlation technique based on word sense merging can be made to yield useful results, given the appropriate tools. Although sense matching across dictionaries in this case too is still prone to errors, there are several reasons why the effort is still worthwhile. Firstly, the proportion of correct sense matches across MRD sources is likely to be high. Secondly, there are many instances in which an incorrect sense-to-sense match does not affect the final result since the information with respect to which a sense correlation is being sought may generalise across closely related word senses. Thirdly, a close inspection of infelicitous matches provides a better understanding of specific difficulties involved in the task and may help us develop better solutions or refine our criteria for sense discrimination.

Sanfilippo & Poznanski (1991) investigated the plausibility of semi-automatic sense correlations with LDOCE and the Longman Lexicon of Contemporary English (LLCE) — a thesaurus which was developed from LDOCE and with which there is substantial overlap (although not identity) between the definitions and entries. Their general goal in developing an environment for correlating MRDs was to provide a Dictionary Correlation Kit containing a set of flexible tools that could be straightforwardly tailored to an individual user's needs, along with a facility for the interactive matching of dictionary entries. Entries are compared along a number of user specified dimensions, such as headword, grammatical code, overlap of the base form of content words in the definition, and so forth, and if an experimentally determined threshold of similarity is found senses are correlated, otherwise the user is asked to make the decision. A trial run with correlation structures derived for 1194 verb senses (over 1/5 of all verb senses in LLCE) yielded encouraging results, with a rate of user interactions of about one for every 8-10 senses and a very low incidence of infelicitous matches (below 1%). We plan further experiments with less closely related sources.

4.2 Mapping from LDB to LKB

4.2.1 Design of the LRL

We chose to use a graph unification based representation language for the LRL, because this offered the flexibility to represent both syntactic and semantic information in a way which could be easily integrated with much current work on unification grammar, parsing and generation. In contrast to DATR (Evans & Gazdar, 1990), for example, the LRL is not specific to lexical representation. This made it much easier to incorporate a parser in the LKB system (for testing lexical entries) and to experiment with notions such as lexical rules and interlingual links between lexical entries. Although this means that the LRL is in a sense too general for its main application, the typing system provides a way of constraining the representations, and the implementation can then be made more efficient by taking advantage of such constraints.

Our typed FS mechanism is based on Carpenter's work on the HPSG formalism (Carpenter 1990, 1991), although there are some significant differences, for example, we augment the formalism with a default inheritance mechanism. This can be used to organise the lexicon in a completely user-defined way, to allow morphological or syntactic information to be concisely specified, for example, as has been done with DATR and other systems. However much of the motivation behind our formalisation of default inheritance comes from consideration of the sense-disambiguated taxonomies semi-automatically derived from MRDs, which we are using to structure the LKB. The top level of the inheritance structure, which is too theory-dependent and abstract to be automatically derived from MRDs, is, in effect, given by the type system. The notion of types, and features appropriate for a given type, gives some of the properties of frame representation languages, and allows us to provide a well-defined, declarative representation, which integrates relatively straightforwardly with much current work on NLP and lexical semantics. However, 'lower' in the hierarchy the nature of lexicographers' classifications forces a default framework (see the next section).

The operations that the LRL supports are (default) inheritance, (default) unification and lexical rule application. It does not support any more general forms of inference and is thus designed specifically to support restricted lexical operations, rather than general reasoning. The type system provides the non-default inheritance mechanism and constrains default inheritance. We use lexical rules as a further means of structuring the lexicon, in a flexible, user definable manner, but lexical rules are also constrained by the type system. The type hierarchy defines a partial order on the types and specifies which types are *consistent*. Only FSs with mutually consistent types can be unified — two types which are unordered in the hierarchy are assumed to be inconsistent unless the user explicitly specifies a common subtype. Unification of FSs is only defined if the meet of their types exists. A full description of the LKB is given in Copestake (1991) and de Paiva (1991).

One advantage of a typed LRL in a large collaborative project is that once an agreed type system is adopted, the compatability of the data collected by each site is guaranteed (there may of course be problems of differing interpretation of types and features but this applies to any representation). In an untyped feature system, typographical errors and so on may go undetected, and debugging a large lexical template based system (see 2.3) can be extremely difficult; a type system makes error detection much simpler. Since a given FS has a type permanently associated with it, it is also much more obvious how information has come to be inherited than if templates are used. The type system can also be integrated with tools for semi-automatic analysis of dictionary definitions, and initial work on this is described in Rodriguez *et al.* (1991). Typing provides the main means of error checking when representing automatically acquired data. Automatic classification of lexical entries by type, according to feature information, can be used to force specification of appropriate information. It remains to be seen whether this restricted LRL will prove adequate for the representation of lexical semantics, but see Briscoe *et al.* (1990) and Copestake & Briscoe (1991) for proposals concerning logical metonymy (see 2.2) and other phenomena which have been taken to require more powerful formalisms.

4.2.2 Creating semantic inheritance hierarchies

We have extended the LRL with default inheritance, based on default unification and restricted by the type system. In outline, FSs may be specified as inheriting information which does not conflict with their own specification from other FSs. Conflicting information is simply ignored. Since the parent FSs may themselves have been defined as inheriting information, inheritance hierarchies can, in effect, be created. We are interested in the use of this mechanism to structure lexical semantic information in particular. In this case, the inheritance hierarchy connects the parts of the lexical entries which contain lexical semantic information, and we derive it from the taxonomies described in 4.1.2.

For example *autobiography* and *dictionary* are found to be below 'book 1 (1)' in the hierarchy semi-automatically derived from LDOCE, and *lexicon* is found below *dictionary*. Part of the lexical semantic structure specified for nouns which are of semantic type **artifact** is the 'telic role' (Pustejovsky 1989a) which indicates the typical purpose of the object. Thus in the feature structure associated with book 1 (1) the telic role is instantiated to the semantics of the appropriate sense of *read* (notated as 'read_L_1_1'). Since this sense of book denotes a physical object, the lexical entry also contains a feature PHYSICAL-STATE. The following fragment of the representation shows the relevant features:

$$\text{book\_L\_1\_1} \qquad \begin{bmatrix} \textbf{lex-noun-sign} \\ \text{ORTH} = \textbf{book} \\ \text{RQS} = \begin{bmatrix} \textbf{artifact\_physical} \\ \text{TELIC} = \begin{bmatrix} \textbf{verb-sem} \\ \text{PRED} = \textbf{read\_L\_1\_1} \end{bmatrix} \\ \text{PHYSICAL-STATE} = \textbf{solid\_a} \end{bmatrix} \end{bmatrix}$$

Lexical entries for *autobiography* and *dictionary* will be automatically defined to inherit their semantic representation by default from book_L_1_1, when their LKB entries are created.

Assuming that no conflicting information is specified for *autobiography* it will have the same values as book_1_1 for both telic and physical-state. However, *dictionary* should be specified as having the TELIC predicate (refer_to_L_0_2) which overrides that inherited from book_L_1_ 1, giving the following partial structure:

$$\text{dictionary} \qquad \begin{bmatrix} \textbf{lex-noun-sign} \\ \text{ORTH} = \textbf{dictionary} \\ \text{RQS} = \begin{bmatrix} \textbf{artifact\_physical} \\ \text{TELIC} = \begin{bmatrix} \textbf{verb-sem} \\ \text{PRED} = \textbf{refer\_to\_L\_0\_2} \end{bmatrix} \\ \text{PHYSICAL-STATE} = \textbf{solid\_a} \end{bmatrix} \end{bmatrix}$$

Since *lexicon* is under *dictionary* it will inherit the more specific value (refer_to_L_0_2) for the telic role.

Attributes such as the telic role are currently being manually associated with entries such as book_L_1_1 which occur as non-leaf nodes in the inheritance hierarchies. This is cost effective since book_L_1_1 directly or indirectly dominates over 100 other entries, and in most cases the default inheritance is not overridden.

Vossen and Copestake(1991) and Vossen (1990b,c) discuss a number of problems which arise in the derivation of inheritance structures from genus taxonomies for Dutch and English nouns; these include, cross-linguistic differences in choice of genus for certain classes, the semantics of conjoined and disjoined genus phrases, weakly classifying genus terms, and the relationship between genus and definiendum.

4.2.3 Productive sense extensions

Lexical rules are essentially just typed FSs which represent a relationship between two lexical entries; the mechanism has been described in Copestake & Briscoe (1991) and I will not discuss it in detail here. Although lexical rule application is typically seen as a way of deriving new lexical entries (but see Krieger & Nerbonne, 1991) our formalisation of lexical rules also allows them to be viewed as a way of describing a relationship between existing lexical entries, and possibly augmenting the information contained in them. Thus lexical rules provide a mechanism for linking lexical entries in a way which can be defined by the user using the type system. One application is for the description of derivational morphology, another is in the representation of sense extensions, such as the use of a word which primarily denotes an animal to denote the flesh of that animal (e.g. *lamb*). By representing sense extensions with lexical rules in the LRL we can allow the relationship between two word senses extracted from an MRD to be described (e.g. lamb 1 (1) and lamb 1 (2) in LDOCE) or, alternatively, fill in gaps in the source data (e.g. deriving a sense meaning 'haddock flesh' from the LDOCE entry for *haddock* which provides only the animal sense). The lexical rule states that a lexical entry of type animal and count noun can be mapped to one of type animal flesh and mass noun. The blocking of lexical rule application is well recognised in morphology (e.g. the existence of *thief* makes the derivation *stealer* unlikely); thus, it is interesting to note that the same phenomenon applies to this type of sense extension – the existence of *pork* makes the use of *pig* to denote pork very marked.

Ostler & Atkins (1991) give further examples of regular extensions of this type which a dictionary, because of its static nature, is forced to simply list (with consequent inevitable errors of omission). However, by mapping the data derived from an MRD source into a LKB supporting lexical operations, it is possible to represent the information in a fashion which does not proliferate sense distinctions, which generalises beyond the source, and which integrates with an account of the parsing process (see Copestake & Briscoe (1991) for further details).

4.2.4 Recognising Verb Classes

Through LDB queries, information is made available which specifies properties of individual word senses (e.g. orthography, pronunciation, part of speech, predicate sense). This information can be semi-automatically integrated with the information structures associated with LKB types by defining a conversion function which establishes correspondences between information derived through LDB queries and values of the features specified by types. For example, Sanfilippo (1991) shows how information for psychological verbs derived through LDB queries can be related to a type system for English verbs. The inheritance network of types represents detailed information about syntactic and semantic properties of verb classes in a form which can be easily tailored to suit requirements of specific NLP systems.

Using the results of the dictionary correlation study briefly described in 4.1.3, LDB queries combining information from LDOCE and LLCE were run which made it possible to individuate members of six subtypes of psychological verbs:

(7)	STIMULUS ARGUMENT	EXPERIENCER ARGUMENT	ILLUSTRATIVE EXAMPLE
	non-causative source	neutral, reactive, emotive	*experience*
	non-causative source	positive, reactive, emotive	*admire*
	non-causative source	negative, reactive, emotive	*fear*
	neutral causative source	neutral, affected, emotive	*interest*
	positive causative source	positive, affected, emotive	*delight*
	negative causative source	negative, affected, emotive	*scare*

These subtypes were defined taking as parameters affect type (positive or negative) and the syntactic realization of the experiencer and stimulus arguments which was semantically characterized as follows:

> Psychological verbs with experiencer subjects are 'non-causative'; the stimulus of these verbs can be considered to be a 'source' to which the experiencer 'reacts emotively'. By contrast, psychological verbs with stimulus involve 'causation'; the stimulus argument may be consided as a 'causative source' by which the experiencer participant is 'emotively affected'. (Sanfilippo, 1991)

The recognition of psychological verbs and their classification into the six subclasses above was facilitated by the inclusion of LLCE set identifiers and by the use of a lexicon from LDOCE containing subcategorisation information (Carroll & Grover, 1989) in LDB queries. For example, whenever a verb sense was associated with the information in (8) a base entry of type, `strict-trans-sign` — the LKB type which describes syntactic and semantic properties of (strict) transitive verbs — was automatically created for that verb sense.

(8) `((Cat V) (Takes NP NP) ...)`

Using this technique, verb types were assigned to some 200 verb senses; this assignment yielded a verb lexicon of 431 entries. The entry below (specified in path notation) provides an illustrative example relative to one of the six semantic varieties of psychological verbs taken into account.

(9)
```
experience L_2_0
STRICT-TRANS-SIGN
< cat : result : result : m-feats : diathesis > = NO-INFO
< cat : result : active : sem : pred > = P-AGT-REACTIVE-EMOTIVE
< cat : result : active : sem : arg2 > = (E-ANIMAL E-HUMAN)
< cat : active : sem : pred > = P-PAT-SOURCE-NO-CAUSE
< cat : active : sem : arg2 > = E-ABSTRACT
< lex-sign sense-id : sense-id dictionary > = "LDOCE"
< lex-sign sense-id : sense-id ldb-entry-no > = "12364"
< lex-sign sense-id : sense-id sense-no > = "0".
```

When loaded into the LKB, (9) will be expanded into a fully-fledged representation for the transitive use of *experience*; by integrating word-specific information provided by (9) with the information encoded by the LKB type `strict-trans-sign`. Thus, although neither LDOCE, LLCE or the earlier subcategorised lexicon contain all the information about psychological verbs defined in Sanfilippo's type system, by using the conjunction of information available from all three, it proved possible to effectively enrich this information at the same time as mapping it into a formal representation.

4.2.5 Towards a Multilingual LKB

A goal of ACQUILEX is to demonstrate that an LKB can be produced that usefully exploits various MRD sources and integrates multilingual information. The use of a common LRL with a common type system, makes it possible to describe lexical entries in a shared 'metalanguage'. This allows lexical entries to be compared, in order to enrich a monolingual LKB or to provide multilingual 'translation links' to create an integrated multilingual LKB. To ensure that information from monolingual sources for different languages can be represented in a compatible way Rodriguez *et al.* (1991) have developed a system for generating LKB entries from analysed definitions which produces entries in the common LRL. In order to do this it is necessary to provide a way of representing equivalences between attribute and value names extracted by the parser, and the feature and type names in the LRL.

Thus we have the beginnings of a semi-automatically constructed LKB containing information from different sources and languages. However, to create a genuinely integrated system, we need to link the senses of the different sources. The general problem with automatically generating links between word senses, either multilingually or monolingually, is that of sense selection. In the case of translation links, we are attempting to link entries derived from two monolingual dictionaries. Bilingual dictionaries, where they do discriminate between senses, will typically use a different scheme from the monolinguals (Van Dale's dictionaries being one exception), and the translations given in bilinguals typically have no sense marking. For example, the English—Spanish VOX dictionary published by Biblograf contains the following entry:

> **crush** *s.* compresión, presión, aplastiamiento, machacadura, estrujamiento, estrujón. *2* cantidad de material machacado, estrujado, etc...

However, not all of these translations will be appropriate for all the senses given in LDOCE for *crush*, and even those which are appropriate may also have inappropriate senses. The approach we take is to attempt to choose word senses which are appropriate translations of a source word sense, by comparison of the information stored in their LKB entries. Copestake & Jones (1991) have developed a general FS matching utility, which can be used to find the best match between a set of candidate LKB entries (where the candidates may have been identified using bilingual MRDs). A statistic is assigned to each potential match, with the magnitude of the statistic proportional to the quality of match. These statistics can then be compared to yield the most likely sense-to-sense mappings.

Although in the simplest (and commonest) cases, we can regard linked lexical entries as translation equivalent, in general we have to allow for such things as different argument ordering, differences in plurality, differences in specificity of reference and 'lexical gaps', where a word sense in one language has to be translated by a phrase in the other. Rather than attempt to generate information about translation equivalence which can be directly used by a particular MT system, we are attempting to describe the relationship between LKB word senses in such a way that the information could be automatically transformed into the lexical component of a variety of MT systems. We represent the cross-linguistic relationships between lexical entries in terms of *tlinks* (for translation link). In general there may be a many-many equivalence between word senses, but each possibility is represented by a single tlink. The tlink mechanism is expressive enough to allow the monolingual information to be augmented with translation specific information, in a variety of ways. As with other aspects of representation in the LKB, the mechanism is very general; it is up to the users of the system to define appropriate types to constrain it.

5 Conclusions

The previous sections have provided a brief and selective introduction to work on the lexicon and selective description of research undertaken within the ACQUILEX project. Space precludes proper discussion of many aspects of lexical research, such as work on spoken language systems. I will conclude by outlining the way in which work on the lexicon can most productively proceed, but first I want to draw attention to two trends which marr slightly the generally positive outlook for lexical research which I hope has emerged from the discussion above.

5.1 (Re)Usability

Mention of reusability has become so common in work on the lexicon recently that no paper in this area can ignore the issue completely. Nevertheless, I believe that the ideas that lie behind the slogan are sufficiently confused that there is a danger that what began as a laudable goal will become counter-productive. It is clearly desirable that if considerable effort is devoted to the development of substantial LDBs and LKBs that the results of this effort should, as far as possible, outlast the latest theoretical fashions and current generation of NLP systems, and should be available generally within the research community. To this end, the development of standards for the interchange of data through the Text Encoding Initiative (e.g. Amsler & Tompa, 1988) and the creation of a repository, in the form of the Consortium for Lexical Research (Wilks & Guthrie, 1990), for the distribution of such material, as well as software tools for creating LDBs, analysing corpora and the like, is to be welcomed as long as such initiatives are not allowed to overshadow the lexical research itself. However, the goal of developing 'generic', 'theory-neutral' or 'polytheoretical' lexical resources is illusory, and potentially harmful if it is interpreted as a reason to avoid theory and concentrate on the creation of large LDBs.

In the discussion above, I have tried to draw a very definite and precise distinction between a LDB and a LKB. In my terms, the latter represents a theory of the lexicon and is, therefore, a body of information couched in a notation – the LRL – which has a formal and explicit syntax and semantics, and which supports lexical operations which perform valid transformations on this information. An LDB, on the other hand, contains information couched in a far looser and varied syntax, typically with an implicit semantics. In terms of both usability and reusability, at least within NLP, it is clear that a LKB will be superior. It is usable precisely because it will be highly theory-dependent and, therefore, make clear and mechanically derivable predictions, and it will be reusable for the same reasons: if we have two LKBs instantiating different theories, one superior in size, the other in theoretical adequacy, it is likely that we will have enough understanding of the relationship between them to carry over the useful information from the former to the latter. Ingria (1988) describes just such a case of the largely automatic creation of a lexicon containing syntactic information for a new parsing system from the lexicon associated with an earlier and less adequate one. The use of the derived lexicon containing syntactic subcategorisation information in the development of a LKB providing a more adequate characterisation of psychological verbs described in 4.2.4 is another example. On the other hand, the relationship between a LDB and a LKB or between two LDBs is likely to be much more difficult to specify, because of the implicit nature of the semantics of the information contained in a LDB. Nevertheless, a LDB is not 'theory-neutral' — any description language constitutes an interpretation of some sort — rather insofar as the semantics of its description language is obscure and ill-understood, a LDB is not so much 'polytheoretical' or 'theory-neutral' as just vague, and thus of diminished utility. None of this should be surprising, given that it has been a commonplace of the philosophy of science for most of this century that all observation and description is theory-laden (e.g. Hanson, 1958). What is perhaps more pertinent is that McNaught (1990) draws similar conclusions in the context of a description of the MULTILEX project, a major goal of which is to define techniques for creating reusable lexical resources. (Re)Usability requires a greater emphasis on theoretical issues, particularly in the area of lexical semantics, not de-emphasis in favour of a large data gathering excercise conducted in a relative theoretical vacuum.

5.2 Lexicography

The predominant role of the bulk of lexicographers in research on the lexicon within NLP and linguistics has been to supply the latter with extremely useful sources of information, whether in printed or machine-readable form. However, these resources have been treated as finished and fixed objects and lexicographers themselves have not, in general, played an active role in theoretical research on the lexicon. Whilst conventional dictionaries tend naturally to be relatively informal and unsystematic and, by the very nature of their organisation, to focus on the individual word rather than generalisations about classes of lexical items, nevertheless they are extremely comprehensive by the standards of contemporary theoretical research. Furthermore, because dictionary making is a commercial activity, publishing houses have considerable resources and experience of managing large lexicographical projects and an invaluable storehouse of lexical knowledge (mostly in the heads of their lexicographers).

The reasons for this comparative lack of collaboration (or at least rather one-sided and static relationship) stem I think from a perception on both sides that the other has little to offer. On our side, there has never been any excuse for this, on that of the publishers, it is more understandable as they are commercial organisations and printed dictionaries represent the product. However, this viewpoint is being rapidly superseded as electronic publishing for human use becomes a reality, and the prospect of commercial markets for lexicons incorporable into NLP systems looms closer. Already there are obvious signs of more active collaboration, most notably in that centred around the electronic publication of the *Oxford English Dictionary* (Tompa, 1986), but also in projects such as the British National Corpus (e.g. Summers, 1990).

5.3 Future Research

There are many challenging theoretical issues, particularly in the area of lexical semantics, which must be addressed before anything like an adequate LKB for most NLP applications can be constructed. However, the research environment made possible by the availability of LDBs should make exploration and development of lexical theories considerably easier — ideas which until recently had to be tested introspectively or by laboriously searching through printed dictionaries can now often be explored in a matter of seconds (see e.g. papers by Carter and Boguraev & Briscoe in Boguraev & Briscoe, 1989). For this reason, it is important that LDBs developed within projects such as ACQUILEX are made widely available within the research community.

The future in computational lexicology and lexicography lies not in further conversion and exploitation of MRDs, but rather in active collaboration with lexicographers and dictionary publishers. Most dictionary publishers urgently need to make the transition from separate projects developing printed dictionaries to cumulative and ongoing development of LDBs. By providing lexicographers with the tools developed within computational lexicography of the type described above, the process of dictionary making will become very much more productive. In addition, at least some lexicographers are already aware of the opportunities created for novel forms of dictionary organisation and presentation by the move away from the printed medium (Atkins, 1991). By making lexicographers aware of the difficulties and problems encountered in the exploitation of MRDs for NLP, it should be possible to ensure that the next generation of LDBs which emerge from publishing houses are far more easily applicable in NLP (allowing us to move on from the relatively ephemeral issues of MRD exploitation.) Most publishers will also need access to computational linguistic tools to assist their lexicographers with corpus analysis. Such tools would not replace the lexicographer, but would be used to provide an initial organisation and selection of corpus data which should improve both quality and productivity. Once again NLP researchers are well placed to provide such tools.

I hope that this rather brief and very selective introduction has provided the reader not familiar with work on the lexicon or computational lexicography with a comprehensible overview and enough references to follow up points of interest. For those working in the field, I hope that my necessarily sketchy and sometimes contentious analysis has provoked thought rather than offence. Quite correctly, the lexicon is now firmly established on the agenda within (computational) linguistic theory and NLP. However, in my opinion the subfield of computational lexicology and lexicography which has burgeoned around the exploitation of MRDs has now served its useful purpose. The future lies in the more illusive but infinitely more rewarding prospect of fruitful and genuine collaboration between linguists, lexicographers and NLP researchers supported and fostered by their respective funding agencies and companies.

Acknowledgements

This work is supported by ESPRIT grant BRA 3030 entitled 'The Acquisition of Lexical Knowledge for Natural Language Processing Systems' to the Universities of Cambridge, Pisa, and Amsterdam, University College Dublin and the Universitat Politecnica de Catalunya, Barcelona. I am grateful to my colleagues on this project for their discussion about many of the issues addressed here and, in particular, to Ann Copestake and Antonio Sanfilippo for their comments, advice and help. The opinions expressed, however, are those of the author alone, as are any remaining errors.

References

Alshawi H(1989) 'Analysing the dictionary definitions' in Boguraev B and Briscoe E J (eds.), *Computational lexicography for natural language processing,* Longman, London, pp.153–169

Alshawi H, Boguraev B and Carter C(1989) 'Placing the dictionary on-line' in Boguraev B and Briscoe E J (eds.), *Computational lexicography for natural language processing,* Longman, London, pp.41–63

Alshawi H, Carter D, Rayner M, Pulman S & Smith A(1991, in press) *The Core Language Engine,* MIT Press, Cambridge, Ma.

Amsler R A(1981) 'A taxonomy for English nouns and verbs', *Proceedings of the 19th ACL,* Stanford, pp.133-138

Amsler R A & Tompa F(1988) 'An SGML-based standard for English monolingual dictionaries', *Proceedings of the Proc. of 4th Conf. of UW Centre for the New OED,* Waterloo, pp.61–79

Atkins B T(1990) 'The dynamic database – a collaborative methodology for developing a large-scale electronic dictionary', *Proceedings of the International Workshop on Electronic Dictionaries,* OISO, Kanagawa, Japan, pp.23–43

Atkins B T(1991, forthcoming) 'Building a lexicon: beware of the dictionary' in Bates L & Weischedel R (eds.), *Challenges of Natural Language Processing,* Cambridge University Press

Atkins B T & Levin B(1990, forthcoming) 'Admitting Impediments' in Zernik U (eds.), *Lexical Acquisition: Using On-Line Resources to Build a Lexicon,* Lawrence Erlbaum, New Jersey

Bloomfield L(1933) *Language,* Allen & Unwin, London

Boguraev B and Briscoe E J (eds)(1989) *Computational lexicography for natural language processing,* Longman, London

Boguraev B K and Levin B(1990) 'Models for lexical knowledge bases', *Proceedings of the 6th Annual Conference of the UW Center for the New OED,* Waterloo, pp.65–78

Boguraev B, Briscoe E J, Carroll J, Copestake A(1991) *Database Models for Computational Lexicography,* Research Report RC 17120, IBM Research Center, Yorktown Heights, New York

Brent M R(1991) 'Automatic acquisition of subcategorization frames from untagged text', *Proceedings of the 29th ACL,* Berkeley, Ca., pp.209–214

Bresnan J & Kanerva J(1989) 'Locative Inversion in Chichewa: A case Study of Factorization in Grammar', *Linguistic Inquiry, vol.21,* 1–50

Bresnan L & Moshi L(1989) 'Object assymetries in comparative Bantu syntax', *Linguistic Inquiry, vol.21,* 147–186

Briscoe E J & Carroll J(1991) *Generalised probabilistic LR parsing of natural language (corpora) with unification-based grammars,* Technical Report No 224, University of Cambridge, Computer Laboratory

Briscoe E J, Copestake A A and Boguraev B K(1990) 'Enjoy the paper: Lexical semantics via lexicology', *Proceedings of the 13th Coling,* Helsinki, pp.42–47

Briscoe E J, Copestake A A & de Paiva V (eds)(1991) *ACQUILEX Workshop on Default Inheritance in the Lexicon,* Technical Report No 234, University of Cambridge, Computer Laboratory

Byrd R, Calzolari N, Chodorow, M, Klavans J, Neff M & Rizk O(1987) 'Tools and methods for computational lexicology', *Computational Linguistics, vol.13.3,* 219–240

Calzolari N(1991) 'Acquiring and representing semantic information in a lexical knowledge base', *Proceedings of the ACL SIGLEX Workshop on Lexical Semantics and Knowledge Representation,* Berkeley, California, pp.188–197

Calzolari N, Peters C & Roventini A(1990) *Computational Model of the dictionary entry,* ACQUILEX Deliverable 1

Calzolari N & Picchi E(1986) 'A Project for Bilingual Lexical Database System', *Proceedings of the Second Annual Conference of the Centre for the New OED,* University of Waterloo, Waterloo, Ontario, pp.79–82

Carpenter R(1990) 'Typed feature structures: Inheritance, (In)equality and Extensionality', *Proceedings of the Workshop on Inheritance in Natural Language Processing,* Tilburg, pp.9–18

Carpenter R(1991, in press) *The Logic of Typed Feature Structures,* Cambridge University Press, Tracts in Theoretical Computer Science

Carroll J and Grover C(1989) 'The derivation of a large computational lexicon for English from LDOCE' in Boguraev B and Briscoe E J (eds.), *Computational lexicography for natural language processing,* Longman, London, pp.117–134

Carroll J(1990) *Lexical Database System: User Manual,* Esprit BRA-3030 ACQUILEX deliverable no. 2.3.3(c)

Carter D(1989) 'Lexical acquisition in the core language engine', *Proceedings of the 4th Eur. ACL,* Manchester, pp.137-144

Chomsky N(1970) 'Remarks on Nominalization' in Jacobs R and Rosenbaum P (eds.), *Readings in English Transformational Grammar,* Ginn, Waltham, Mass.

Church K(1985) 'Stress assignment in letter-to-sound rules for speech synthesis', *Proceedings of the 23rd ACL,* Chicago, Illinois, pp.246–254

Church K and Hanks P(1990) 'Word Association Norms, Mutual Information and Lexicography', *Computational Linguistics, vol.16,* 1

Clear J(1987) 'Computing' in Sinclair J (eds.), *Looking up: An Account of the COBUILD project in Lexical Computing,* Collins ELT, London and Glasgow

Copestake A A(1990) 'An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary', *Proceedings of the Workshop on Inheritance in Natural Language Processing,* Tilburg, pp.19–29

Copestake A A(1991) 'The LKB: a system for representing lexical information extracted from machine-readable dictionaries', *Proceedings of the ACQUILEX Workshop on Default Inheritance in the Lexicon,* Cambridge

Copestake A A and Briscoe E J(1991) 'Lexical Operations in a Unification Based Framework', *Proceedings of the ACL SIGLEX Workshop on Lexical Semantics and Knowledge Representation,* Berkeley, California, pp.88-101

Copestake A A and Jones B(1991) *Support for multi-lingual lexicons in the LKB system,* ms University of Cambridge, Computer Laboratory

Evans R and Gazdar G (eds)(1990) *The DATR papers,* Cognitive Science Research Paper CSRP 139, School of Cognitive and Computing Sciences, University of Sussex

Fillmore C J and Atkins B T(1991, forthcoming) 'Risk: the Challenge of Corpus Lexicography' in Zampolli and Atkins (eds.), *Automating the Lexicon II,* Oxford University Press

Flickinger D, Pollard C & Wasow T(1985) 'Structure sharing in lexical representations', *Proceedings of the 23rd ACL,* Chicago, pp.262–267

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

Grimshaw J(1990) *Argument Structure,* MIT Press, Cambridge, Ma.

Gross M(1984) 'Lexicon-Grammar and the syntactic analysis of French', *Proceedings of the 10th Coling,* Stanford, Ca., pp.275–282

Guthrie L, Slator B M, Wilks Y and Bruce R(1990) 'Is there content in empty heads?', *Proceedings of the 13th Coling,* Helsinki, pp.138–143

Hindle D & Rooth M(1991) 'Structural ambiguity and lexical relations', *Proceedings of the 29th ACL,* Berkeley, Ca., pp.229–236

Hobbs J, Croft W, Davies T, Edwards D & Laws K(1987) 'Commonsense metaphysics and lexical semantics', *Computational Linguistics, vol.13,* 241-250

Ingria B(1988, in press) 'Lexical information for parsing systems: points of convergence and divergence' in Walker D, Zampolli A, Calzolari N (eds.), *Automating the Lexicon: Research and Practice in a Multilingual Environment,* Cambridge University Press, Cambridge

Kasper R T & Rounds W C(1990) 'The logic of unification in grammar', *Linguistics & Philosophy, vol.13.1,* 35-58

Kay M(1984) 'Functional unification grammar: a formalism for machine translation', *Proceedings of the 10th International Congress on Computational Linguistics (Coling84),* Stanford, California, pp.75–9

Klavans J L and Wacholder N(1990) 'From Dictionary to Knowledge Base via Taxonomy', *Proceedings of the 6th annual conference of the Waterloo Centre for the New OED and Text Retrieval,* Waterloo

Krieger H and Nerbonne J(1991) ‘Feature-Based Inheritance Networks for Computational Lexicons’, *Proceedings of the ACQUILEX Workshop on Default Inheritance in the Lexicon,* Cambridge

Levin B(1988, in press, 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

Levin B(1990, in press) *Towards a Lexical Organisation of English Verbs,* University of Chicago Press

Liberman M(June, 1991) *The ACL Data Collection Initiative,* Ms. University of Pennsylvania

McNaught J(1990) ‘Reusability of Lexical and Terminological Resources; Steps towards Independence’, *Proceedings of the International Workshop on Electronic Dictionaries,* OISO, Kanagawa, Japan, pp.97–107

Moens M, Calder J, Klein E, Reape M & Zeevat H(1989) ‘Expressing generalizations in unification-based formalisms’, *Proceedings of the 4th Eur. ACL,* Manchester, pp.174–181

Moortgat M, Hoekstra T, van der Hulst H(1980) *Lexical Grammar,* Foris, Dordrecht

Moshier M D and Rounds W C(1987) ‘A logic for partially specified data structures’, *Proceedings of the 14th ACM Symposium on the Principles of Programming Languages,* , pp.156–167

Normier B and Nossin M(1990) ‘GENELEX Project: EUREKA for Linguistic Engineering’, *Proceedings of the International Workshop on Electronic Dictionaries,* OISO, Kanagawa, Japan, pp.63–70

Ostler N and Atkins B T S(1991) ‘Predictable Meaning Shift: Some Linguistic Properties of Lexical Implication Rules’, *Proceedings of the ACL SIGLEX Workshop on Lexical Semantics and Knowledge Representation,* Berkeley, California, pp.76–87

de Paiva V(1991) ‘Types and Constraints in the LKB’, *Proceedings of the ACQUILEX Workshop on Default Inheritance in the Lexicon,* Cambridge

Pollard C(1984) *Generalized Phrase Structure Grammars, Head Grammars, and Natural Language,* Unpublished PhD Dissertation, Stanford University

Pollard C, Sag I(1987) *Head-Driven Phrase Structure Grammar,* University of Chicago Press

Hanson N R(1958) *Patterns of Discovery,* Cambridge University Press

Procter P(1978) *Longman Dictionary of Contemporary English,* Longman, England

Pustejovsky J(1989a, in press) ‘The Generative Lexicon’, *Computational Linguistics, vol.17.3,*

Pustejovsky J(1989b) ‘Current issues in computational lexical semantics’, *Proceedings of the 4th European ACL,* Manchester, pp.xvii–xxv

Rodriguez H et al(1991) *Guide to the extraction and conversion of taxonomies,* ACQUILEX project draft user manual, Universitat Politechnica de Catalunya, Barcelona

de Rose(1988) ‘Grammatical category disambiguation by statistical optimisation’, *Computational Linguistics, vol.14.1,* 31–39

Russell G, Carroll J and Warwick-Armstrong S(1991) ‘Multiple default inheritance in a unification based lexicon’, *Proceedings of the 29th ACL,* Berkeley, pp.215–221

Sager N(1981) *Natural Language Processing,* Addison-Wesley, Reading, Mass.

Sanfilippo A(1990) *Grammatical Relations, Thematic Roles and Verb Semantics,* PhD Dissertation, University of Edinburgh

Sanfilippo A(1991) ‘LKB Encoding of Lexical Knowledge from Machine-Readable Dictionaries’, *Proceedings of the ACQUILEX Workshop on Default Inheritance in the Lexicon,* Cambridge

Shieber S(1984) ‘The design of a computer language for linguistic information’, *Proceedings of the COLING84,* Stanford, California, pp.362–366

Shieber S(1986) *An Introduction to Unification-based Approaches to Grammar,* University of Chicago Press, Chicago

Sinclair J(1987) *Looking up: An Account of the COBUILD project in Lexical Computing,* Collins ELT, London and Glasgow

Steedman M(1985) ‘Dependency and coordination in the grammar of Dutch and English’, *Language, vol.61,* 523-568

Summers D(1990) ‘Longman computerization initiatives’, *Proceedings of the International Workshop on Electronic Dictionaries,* OISO, Kanagawa, Japan, pp.141–152

Tompa F(1986, Unpublished Ms.) *Database Design for a Dictionary of the Future,* University of Waterloo

Uchida H(1990) 'Electronic Dictionary', *Proceedings of the International Workshop on Electronic Dictionaries,* OISO, Kanagawa, Japan, pp.23–43

Vossen P(1990a) *A Parser-Grammar for the Meaning Descriptions of LDOCE,* Links Project Technical Report 300-169-007, Amsterdam University

Vossen P(1990b, forthcoming) 'Polysemy and vagueness of meaning descriptions in the Longman dictionary of contemporary English' in Svartvik J and Wekker H (eds.), *Topics in English Linguistics,* Mouton de Gruyter, Amsterdam

Vossen P(1990c) 'The end of the chain: Where does decomposition of lexical knowledge lead us eventually?', *Proceedings of the 4th conference of Functional Grammar,* Copenhagen

Vossen P & Copestake A(1991) 'Untangling definition structure into knowledge representation', *Proceedings of the ACQUILEX Workshop on Default Inheritance in the Lexicon,* Cambridge

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

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

Wilks Y, Fass D, Guo C-M, McDonald J, Plate T and Slator B(1989) 'A tractable machine dictionary as a resource for computational semantics' in Boguraev B and Briscoe E J (eds.), *Computational lexicography for natural language processing,* Longman, London, pp.193–231

Wilks Y & Guthrie L(1990) 'The Consortium for Lexical Research', *Proceedings of the International Workshop on Electronic Dictionaries,* OISO, Kanagawa, Japan, pp.179–180

Zeevat H, Klein E & Calder J(1987) 'An introduction to unification categorial grammar', *Proceedings of the Edinburgh Working Papers in Cognitive Science, Vol 1,* Categorial Grammar, Unification Grammar and Parsing