FINAL REPORT

ESPRIT BASIC RESEARCH ACTION No. 3030

ACQUILEX

Acquisition of Lexical Knowledge for Natural Language Processing Systems

Date: 8 June 1992

Editors:
T.Briscoe (University of Cambridge - U.K.)
N.Calzolari, A.Zampolli (Pisa University/Institute of Computational Linguistics, Pisa,I)
A.Cater (University College - Dublin - IRL)
W.Meijs, P.Vossen (University of Amsterdam - NL)
F.Verdejo (University of Barcelona - SP)

1. EXECUTIVE SUMMARY

Overview of the work performed during the project

The activities performed during the project can be roughly subdivided into two major phases:

1) The first year was mainly devoted to the production of the tools, the acquisition and definition of the data, and the design of the basic methodologies of work. In the first year of the project, the partners mainly concentrated their efforts, according to the workplan agreed with the EEC in the technical annex (TA), in organizing cooperation among the partners, in the acquisition of lexical sources, and in the creation of the software tools for managing and accessing the lexical data.
The deliverables envisaged by the TA and produced during the first year must be considered primarily in their function as working tools for the execution of the types of research envisaged in the technical annex for the second and third year of the project.

2) One of the results of the research has been to consolidate these tools, through the experience gained in actual use, and to evaluate the possibility and cost-effectiveness of generalizing them so that they can also be used outside the project. The second phase of the project was in fact dedicated to: i) the actual development of the new results (mainly the semantic taxonomies, the Type system, the LKB, its population for the 4 languages of the project (for the semantic subset agreed on), the NLP testbed), and ii) a rather large technical/scientific production of many papers (presented at all the major conferences in Computational Linguistics) and reports, where the main problems - raised by this more creative phase of the work - have been illustrated, tackled and, where possible, solved.

The description of the activities performed during the project, which includes the major milestones achieved, the deliverables produced, the corrective actions taken and problems encountered, is here organized around 10 major Work Parts (WP1 - 10), which are logically concatenated in the flux of the work. At the beginning of this section (part A) organizational matters are described; at the end of the section some concluding remarks are given (part B). WP1 to 10 are structured as follows:

- a brief statement on the nature of the task
- the deliverable(s) concerned
- the partner(s) responsible for reporting on the activities
- the status with respect to the forecasts of the TA
- other information (problems, corrective action, external contacts, impact, etc.).

A. Establishing an adequate level of organization and interaction among the partners

In the first few months we had to overcome organizational problems specific to the very nature of the project: different sites have different experiences, different tools, different infra-structures and different dictionaries. Before reaching full cooperation there was a great deal of communication between the sites.
Furthermore, some tasks, such as conversion of the source dictionaries, had to begin right away before the computational model was fully defined and before the LDB

software was made available to all sites. All sites, therefore, started working on conversion using their own tools and their own approach.

Meetings

In the first year two general meetings were held in Pisa and one was held in Amsterdam to discuss the content and the planning of the first year of the work, to distribute the tasks, and to agree, where needed, on common specifications.

In the second and in the last year, general meetings were held in Cambridge (where a major revision of the Workplan was agreed upon among the partners), two in Barcelona, one in Dublin, and one in Amsterdam.

Many other restricted meetings were held on specific topics to agree on specific actions.

An international workshop on default inheritance was held in Cambridge, and the proceedings will be published by CUP.

Responsibility for the Work Parts

The following distribution of responsibility has been agreed on for the coordination of the activity of the Work Parts and for the preparation and presentation of the Final Report:

WP1: Computational Model of a Dictionary;
Resp: Pisa

WP2: Lexical Requirements of NLP Systems;
Resp: Dublin

WP3: Lexical Database Software and Documentation;
Resp: Cambridge, Pisa

WP4: Conversion of MRD Sources;
Resp: Amsterdam

WP5: Initial Definition of the Vocabulary Subset and Lexical Information for Further Work;
Resp: Amsterdam

WP6: NLP Testbed System;
Resp: Dublin

WP7: Lexical Knowledge Base System;
Resp: Cambridge

WP8: Semantic Taxonomies;
Resp: Pisa

WP9: Evaluation of LDB/LKB Systems;

Resp: Barcelona

WP10: Report on Feasibility of MRD Sources for NLP Systems;
Resp: Amsterdam

A. 1. Organisational Problems: Addition of new partners

At the very beginning of the project an unforeseen amount of time and energy was spent solving a problem which turned out to be more complicated than expected. The Commission suggested to the four original partners (Amsterdam, Cambridge, Dublin, Pisa) the inclusion of two new partners in the project. The discussion among the partners continued for several months, and was cause for some disagreement. It was eventually agreed to include the group of the "Universitat Politecnica de Cataluna" as a fifth partner. The inclusion of Barcelona necessitated an additional contract.

During the second year another partner, Cambridge University Press, had been proposed to the EC by the project partners for inclusion, and was accepted by the EC. An amendment to the original contract was signed for this purpose.

The addition of an industrial partner, in particular a publishing house, should be considered a major achievement for the project; it shows the interest of the publishing community in the tools and methodologies being developed within ACQUILEX.

A.2. Equipment

At the beginning, a certain amount of time was spent, by some groups, for the acquisition of the hardware and software necessary to execute the work, and in particular to ensure the common utilisation of the software produced in Cambridge.

The Pisa group has:
- made available 6 workstations in MS/DOS and one workstation in UNIX;
- set up a NOVELL network to link the different workstations involved in the project;
- a connection to the IBM mainframe of CNUCE, with a VM operating system.

The Cambridge group has:
- an APPLE Macintosh II ci
- a DEC MIPS (Unix workstation).

The Amsterdam group has:
- three Macintoshes:
  II cx: 8 Mb internal memory and 80 Mb harddisk
  SE30: 4 Mb internal, 40 Mb harddisk
  PLUS ED: 1 Mb internal, 20 Mb harddisk
  Separate 240 Mb external harddisk
- a Tops Network relating the Macintoshes.
- a connection to the VAX mainframe of the Faculty of Computational Linguistics, with a VMS operating system.

The Dublin Group has:

- one Mac II cx, 8 Mb and 80 Mb disk, linked by Ethernet to other machines of all types.

The Barcellona group has:
- one Mac II-ci: 8 Mb internal memory, 80 Mb hardisk
- a connection to the VAXes network of the Technical University of Catalunya with an VMS operating system.

A.3. Acquiring the rights for using machine-readable dictionaries.

Cooperation with publishing houses

The availability of machine-readable dictionaries was obviously a condition sine qua non for the feasibility of the project. In addition, during the negotiations for the acquisition of the dictionaries, it appeared that the issue of the cooperation with publishing houses was an important aspect of the problem connected to the creation of reusable lexical resources for the NLP community.

There are at least two different, although possibly related, ways in which lexicographers can be involved in the creation of lexical databases:

- A publishing house gives permission to use its MRDs to extract lexical information to be incorporated in the LDB.

- A publishing house can cooperate in the creation of LDBs by offering the know-how and the work of its lexicographers.

There may be several reasons for the decision of lexicographers to cooperate:

- to influence the design of a lexicographic workstation and to share software components for lexicographical works;

- to use the resulting LDB as a source from which to derive, in a computer-assisted manner, various types of diversified lexicographical products;

- to distribute new types of electronic dictionaries, both for the human users, and as a component of some NLP systems (e.g. lexica, synonym lists, thesauri for spelling checkers and information retrieval).

In either case, the copyright issue is becoming an increasingly difficult problem to solve. Interest for the cooperation between NLP and publishing houses has increased rapidly over the past few years, mainly on the part of NLP but also on the part of some publishing houses. The surge of initiatives in the sector of textual corpora has added to the surge of initiatives for dictionaries.

The problem is complicated by: i) differences in national legislation; ii) the difference and novelty of the uses of the copyright materials within NLP, with respect to the traditional ones; iii) the diverse levels of understanding of these uses as shown by different publishing houses; iv) a different grade and type of interest by the publishing houses for the potential of language industries.

In recent discussions, it became evident that there is an urgent need to outline a precise typology of the various utilisations of copyrighted material, and to formulate guidelines for the copyright agreements with the publishing houses. Actions to be taken in the copyright field have been suggested to the EEC, who in turn requested that the coordinator of the project formulate a concrete proposal through a feasibility study. A report on similar issues concerning textual corpora will be produced within the NERC project.

While awaiting further coordinated developments at the European level (EEC and Council of Europe), each partner has independently acquired the right to use the relevant dictionaries. It should be noted that, in addition to the use of dictionaries, other forms of cooperation have been envisaged with several publishing houses. The Pisa Group is discussing an agreement with OUP for cooperation in the field of meaning analysis. The Amsterdam Group has come to an agreement with Van Dale publishers. The Cambridge Group has contacts with OUP, Longman, and CUP regarding dictionary development.

The importance, and the feasibility, of involving publishing houses in research projects like ACQUILEX has found its realization in the composition of the consortium which will be involved in the follow-up of ACQUILEX, i.e. ACQUILEX-II, where three publishing houses have joined the original partners: CUP (already present in the second phase of this project), Van Dale (Amsterdam) and Vox (Barcelona).

A.4. Dictionaries available within the Project

The dictionaries used by the different groups are the following:

Pisa
Il Nuovo Dizionario Italiano Garzanti, Garzanti, Milano, 1984.
Collins Concise English-Italian, Italian-English Dictionary, Collins, London and Glasgow, 1985.
DMI, *Dizionario di Macchina dell'Italiano*, based on the Zingarelli Italian Dictionary, ILC.

Amsterdam
W. Martin and G.A.J. Tops, *Groot Woordenboek Engels-Nederlands*, Van Dale Lexicografie, Utrecht, 1984.
W. Martin and G.A.J. Tops, *Groot Woordenboek Nederlands-Engels*, Van Dale Lexicografie, Utrecht, 1986.
P.van Sterkenburg and W.J.J. Pijnenburg, *Groot Woordenboek van Hedendaags Nederlands*, Van Dale Lexicografie, Utrecht, 1984.

Cambridge
Longman Dictionary of Contemporary English, Longman, London, 1978.
Longman Lexicon, Longman London, 1981.
Collins COBUILD, Collins, London, 1987.
Oxford Advanced Learners Dictionary, OUP, 1974.

Barcelona
Vox: *Diccionario Illustrado de la lingua Espanola*, Biblograf, 1987.

WP1. COMPUTATIONAL MODEL OF A DICTIONARY

Goal

The goal is to produce a working definition of the computational model of a dictionary, specifying a general representation for a lexical entry capable of handling a variety of different MRD sources and derived lexicons.

Deliverables

The report on the model constitutes a 6-month deliverable:
N. Calzolari, C. Peters, A. Roventini, "Computational Model of the Dictionary Entry", Preliminary Report, Pisa 1990. (Deliverable No. 1)

At 30 months an evaluation of the preliminary report has been performed. This has been formulated in a brief report containing the Final Proposal for the project Common Lexical Entry.

Responsibility

The Pisa group is responsibile for this deliverable, with the cooperation of Amsterdam, Barcelona, Cambridge.

Status with reference to the Workplan: (on time)

Before the start of the project, each partner had already independently acquired experience in developing procedures to decode and interpret some MRDs and to store their content in some database format.

At the beginning of the project it was felt necessary to establish a common ground where it could become feasible to communicate among the partners about the MRD data thus avoiding misinterpretation, and to exchange or to share data and results. Even more importantly, the need of storing all the different MRD sources into the same LDB system required a common method of representing and interpreting MRD data.

The steps by which this task has been carried out are the following:

- A detailed analysis has been performed of the structure and content of all the project dictionaries. This part of the work has been developed in each site for the 'local' MRDs.

- A proposal has been made concerning the method of representing MRD data in a uniform way by using:
  - a 'common but not fixed' template structure;
  - a common set of Node-tags and Attribute-tags;
  - lists of the possible values for the Attribute-tags.

- While applying this method to all the MRDs of the project, revisions and modifications appeared necessary (also to adjust features characteristic of each

dictionary). In its present form, the model has been used for a uniform description and representation of all the MRD sources available to the project.

- The preliminary format for the Common Lexical Entry (CLE) was discussed and decided upon at the Amsterdam meeting. Its main purpose was for building the common Lexical Knowledge Base (LKB) and for the NLP tasks of the testbed system (for whose application it was judged useful to have word-sense based entries).

All the dictionaries have been converted into this CLE model to deal with the tasks for which it was conceived.

Usability outside of the project

We do not have to exchange the MRDs proper, but the already 'analysed' MRD data; we therefore need more a model and an 'explicit' method for a uniform representation and interpretation of analysed data than a formalism for a uniform encoding of the source data.

This model can be used as a starting point for a fairly straightforward translation in a formal language, e.g. in SGML, which coincides with the objectives of the TEI (Text Encoding Initiative), to which our model has been supplied. In fact, our model has been used within the Printed Dictionary Work Group of the TEI as one of the bases of the TEI proposal for encoding MRDs.

This model has also been used and/or evaluated by other research groups working on MRDs (e.g. in Prague, at Stuttgart University) and is judged to be easily mappable and compatible with other models (e.g. the IBM Hawthorne).

An abridged version has been incorporated, as a proposal to be considered in the feasibility study for reusing lexical resources, in the Survey part of EUROTRA-7.

WP2. REPORT ON THE LEXICAL REQUIREMENTS OF NLP SYSTEMS

Identification of the various classes of lexical requirements in various types of NLP tasks

Goal

The goal was to identify the various types of information that a lexicon should provide to other linguistic components in NLP systems, and to exemplify the classes of related linguistic processing problems. The report provides criteria for the development of the LDB/LKB, and for the definition of the information to be extracted for the case studies.

Deliverables

A six-month report: A. Cater, C. Guo, "Lexical Requirements of Natural Language Processing", Dublin, July 1990. (Deliverable No. 2).

Responsibility

Dublin

Status with reference to the TA

The deliverable was released with a five-month delay, due to bureaucratic delays in the signing of the contract.

Impact on other WPs

Dublin preferred the option of defining the various types of information at a very general level as opposed to the alternative approach of making a detailed description of the lexical information used in a variety of existing NLP system.
The report discusses, in particular, both i) lexical types of knowledge for the acquisition and representation for which the present state-of-the-art provides adequate know-how, and ii) types of knowledge whose nature, acquisition, and representation are still problematic andthat are the object of advanced research still underway.

In the second part of the project, the information gathered in this WP has been used, in cooperation with Dublin and the groups working on the converted dictionaries, for evaluating the relevance of the information which can be extracted from available dictionaries, defining priorities in the choice of case studies, and ensuring the reusability of the knowledge extracted from the dictionaries in the NLP components created in Dublin.

WP3. LEXICAL DATABASE SOFTWARE AND DOCUMENTATION

Designing, creating and testing software tools for interactive access of the LDB

Goal

The goal was to create a system which provides facilities for loading, querying and extracting information from the MRDs loaded in the LDBs of the project.

Deliverables

The TA foresees two deliverables:

- at six months: First release of LDB software and documentation. A technical description is given in J.Carroll, "Lexical Database System User Manual", Cambridge, April 1990. (Deliverable No.3).

- at twelve months: Second release and demonstration at the workshop. A technical description is given in J.Carroll, "Lexical Database System User Manual", Cambridge, October 1990, and E.Marinai, C.Peters, E.Picchi, "The Pisa Multi-Lexical Database System", Pisa, November 1990. (Deliverable No.4).

Responsibility

The Cambridge and Pisa groups have worked on the LDB software. Instead of choosing a common implementation language, the Technical Annex has assigned:
- to Cambridge, the implementation in Common Lisp of the common software, to be installed, tested and used by all partners;
- to Pisa, the implementation in a MS/DOS context, using Pascal and Prolog.

The reasons behind the decision were the following:

- to allow two different approaches, based on different existing experiences and expertise, to be developed and compared;

- to cover a wide range of computer hardware;

- to connect the LDB software with a variety of external applications (for the Cambridge group, various activities on parsers in an AI/NLP framework; for the Pisa Group, a wide network of human users of various disciplines using lexical resources for computer-aided access to large textual corpora).

Status with reference to the TA: (on time)

The 12-month release of the LDB incorporates, as foreseen by the TA, facilities for the extraction, comparison and integration of information from the different LDB sources. A user-interface supports simultaneous access to the various MRDs loaded into the system. Specialized pattern-matching tools allow extraction of information from specific

fields of the dictionaries. Merging of this information to create new entries and derived lexicons is possible.

WP4. CONVERSION OF MRD SOURCES

Creation of the necessary conversion software tools for decoding the dictionary tapes and transferring the data in the required formats

Two alternative approaches have been considered and discussed:

a) An attempt to design and to implement a common general software, generalized at least over the various MRDs involved in the project, and possibly further generalizable in order to create a reusable MRD parser for the research community.

b) Designing and writing specific software tools for parsing and converting each dictionary (or a homogeneous group of dictionaries).

The second alternative has been adopted for the reasons cited below:

- producing generalized software, not foreseen in the original TA, would have required so much time that the effective performance of the conversion would have been delayed, hence delaying the entire project, which depends completely on having the MRD data available;

- the conversion procedure is very time-consuming, and specific software will optimize the execution time;

- results and experiences acquired in other research contexts have shown that, although an interesting challenge at the intellectual level, the creation and the use of a generalized parser is not economically viable.

Performing the actual conversion of the lexical sources

Goal

Parsing the photocomposition tapes of all the dictionaries available within the Project, and transferring the data into the required format.

Deliverables

The dictionaries converted into the required format constitute, in the TA, a deliverable for the twelve months of the project. A brief description of the process performed is given in "Conversion of Machine Readable Dictionary Sources", Amsterdam, November 1990 (deliverable No. 5).

Responsibility

Amsterdam coordinates the reporting on this work part. Amsterdam, Barcelona, Cambridge, and Pisa were involved in this WP.

Status with reference to the TA: ***(on time)***

The time and cost required for this operation obviously vary from case to case according to the degree of consistency, univocity and formalization of the encoding system, which is reflected in the number of manual interventions necessary to support the processing through the decoding software. All the dictionaries acquired have been converted.

Comments

It should be observed that, on the whole, these dictionaries probably form the most extended lexical resource which has been created so far and, with the limitation of the different copyright agreements, they are a potential patrimony, not only for any follow-up of the project, but for a large number of R&D activities in the field of language engineering and language industries.

WP5. INITIAL DEFINITION OF VOCABULARY SUBSET AND LEXICAL INFORMATION FOR FURTHER WORK

Exploring and assessing possible criteria for the definition of a common lexical subset, to be used in further research during the second phase of the project

Goal

The goal was to identify test cases and to define an appropriate (about 2,000 words-sense) vocabulary subset. This subset should constitute the lexicon for the testbed analyzers and generators.

Responsibility

Amsterdam has coordinated the discussion and is responsible for the presentation of the reports.

Deliverables

A 12-month report: "Initial Definition of the Vocabulary Subset", Amsterdam, November 1990. (Deliverable No. 6).

Status with reference to the TA

The original TA has been slightly modified in order to allow for the final choice of the vocabulary subset on the basis of information acquired through the anticipation of some other part of the project (in particular, taxonomy building).

Different approaches were advocated to decide upon the appropriate choice. Moreover, the various groups differed in their interpretation of the TA requirements. Some felt that the 12-month deliverables should already include the list of the 2,000 word-senses which should constitute the lexical subset. Others felt that only the selection criteria were to be defined, and that it was necessary to progressively perform the choice of the actual word-senses in parallel with the operations for the creation of the LKB.

In the end, the second alternative was adopted. The main reason for doing so was the need to initially analyze the dictionaries with our processing tools in order to ensure the maximal representativity of the final sample. It is necessary to explore various syntactically and semantically defined classifications of the lexicon and their relationship with specific classes of NLP problems in order to choose an appropriate set of words which, although limited in number, cover a large variety of lexical types, of phenomena tackled for NLP tasks, and of methods of extraction of information.

Therefore, in the initial state of the deliverable, a range of criteria were identified, and afterwards a process of experimentation and evaluation was begun.

Corrective Actions

The central concern is to ensure that the composition of the lexical subset:
- represents the various relevant lexical classes and their distribution;
- provides lexical information for the parser and the generators according to their needs and according to the points they wish to demonstrate.

Therefore, the choice of the word-senses to be included in the lexical subset must:
- utilise the different classifications of the lexicon, generated on the basis of the different semantic relations progressively extracted from the definitions;
- interact with the results in the use of the parser and generators prototypes, as they gradually become available.

Consequently, part of the resources and time foreseen for this task during the first year were dedicated to anticipating the work regarding the retrieval of the semantic relations and the analysis of specific taxonomies (see WP 8).

WP6. THE NLP TESTBED SYSTEM

Goal

To develop an English sentence analyser and sentence generators for Dutch and Italian. The purpose is not to produce a practical system, but to produce a testbed for evaluation of the relevance for NLP tasks and the quality of the lexical information extracted from MRDs. To provide the minimal functionality required by the testbed, the NLP system will analyse sentences using "deep" conceptual representation and will generate Dutch and Italian sentences from the same representation.

Deliverable

The 12 months deliverable is represented by the first release of the testbed English language analyser. A technical description is given in A. Cater, "Testbed English Language Analyser", Dublin, November 1990. (Deliverable No. 7).

Given the agreement with the EC of not having the 2nd year Review, the other reports and deliverables foreseen in the TA for this WP have been brought together in a unique Deliverable. (The second part of the deliverable will be presented in a final version at the end of July).
A. Cater, P. Matthews, "NLP Testbed and LKB", Dublin, June 1992. (Deliverable No. 10).

Responsibility

Dublin, with data provided by all the other partners.

Status with reference to the TA

Workpart 6 began in December 1989 with the appointment of a senior research assistant, Dr. Chengming Guo, who began to work with Dr. Arthur Cater on Deliverable No.2 "Lexical Requirements of Natural Language Processing systems". In April 1990 a junior assistant, Mr. Paddy Matthews, was also appointed. This first report was made available in August 1990. It assisted in concentrating the project team's attention on those aspects of lexical knowledge whose extraction from human-oriented dictionaries was especially important for practical NLP applications, and it also identified important kinds of knowledge which would not be found in dictionaries.

In mid 1990 also, work began on modifying an existing English sentence analyser, with a view to bringing the structure of its lexicon more into line with the Computational Model of the Lexicon which was emerging from other parts of the project. The sentence analyser, forming a crucial part of the overall testbed system, was presented in Deliverable No.7, and the software was demonstrated at the first project review in November 1990. Its particular emphasis on the importance of non-syntactic lexical information were cause for considerable discussion among some who were working with dictionaries, who feared that this sort of information could not be extracted.

In January 1991 a two-day workshop was held, attended by representatives from almost all sites. This workshop refined the goals of the entire testbed workpart, establishing a range of attainable and lexically significant syntactic phenomena that the Dutch and Italian translation components should cover.

By November 1991 the sentence generators for Dutch and Italian, though not ready for demonstration, were almost complete. Meanwhile, work had begun on using lexical information derived from the LKB to drive the English sentence analyser. Preliminary results from this work were presented at a project workshop in November 1991. These results prompted a minor reappraisal of the approach to representing closely related usages of English verb senses. They also gave the first concrete indication that one of the projects major evaluative criteria was being met: reusability of lexical resources in different theoretical frameworks.

At the end of the project, similar work was still under way on exploiting LKB information for the purposes of the generators. Unfortunately a considerable amount of hand-coding of information is having to be performed, most notably because verbs are only available in English in the LKB at present, because data on verbs have been derived from the Longman lexicon. Work towards their extraction from available dictionaries is still going on in Barcelona and Pisa. Nevertheless, work on the generation of English is also being carried out in an effort to demonstrate that the LKB will provide reusable information that is of utility in generation as well as analysis of language.

Contacts with external parties

The "Testbed English Language Analyser" software is a further development of software that was developed as part of a research contract awarded to Dublin's Artificial Intelligence Research Centre by Digital Equipment Corporation. It is not, however, anticipated that future modifications undertaken as part of ACQUILEX, in particular those modifications concerned with integration with the LDB software, will be communicated to Digital or indeed to any other party.

WP7. LEXICAL KNOWLEDGE BASE SYSTEM

Goal

Development of knowledge representation language to support lexical information.

Deliverable

Given the agreement with the EC of not having the 2nd year Review, the description of this WP has been brought together in a unique final Deliverable, but the LKB system was released at the 18 and 24 month points as foreseen in the TA.
T. Briscoe, A. Copestake, "LKB", Cambridge, May 1992 (Deliverable No. 8).

Responsibility

Cambridge is responsible for producing the software.

Status with reference to the TA: (on time)

The ACQUILEX lexical knowledge base (LKB) system has been designed to allow the representation of syntactic and semantic information extracted from MRDs. The LKB implements the representation of reusable multilingual information in a structured lexical knowledge base in a unification-based typed feature structure language which supports lexical operations such as (default) inheritance and lexical rule application.

Substantial monolingual lexicon fragments have been derived (semi-)automatically for English, Spanish, Italian and Dutch and represented in the LKB, and cross-linguistic links between these entries have been generated semi-automatically. Working paper (AWP) 036 (Copestake, 1992a) is a relatively general paper describing the LKB software and aspects of its use within ACQUILEX.

WP8. SEMANTIC TAXONOMIES

Goal

The goal was the extraction from all the project monolingual MRDs of semantic taxonomies for the agreed vocabulary subset. These taxonomies constitute the basis for building the Lexical Knowledge Base (LKB). Under this WP we also consider the successive, and more demanding, step consisting in the extraction of other semantic information from the *differentia* part of the definitions, to be represented in the Type System of the LKB.

Deliverable

Production of semantic taxonomies for the agreed vocabulary subset is an 18 month deliverable.
N. Calzolari, T. Marti, P. Vossen, "Taxonomies and Feature Structures", Pisa, November 1991 (Deliverable No. 9).

Responsibility

Pisa is responsible for reporting on this work part, with the contribution of Amsterdam and Barcelona.

Status with reference to the TA: (on time)

Part of the work has been anticipated, as a consequence of the revision of the TA.

All of the groups considered it useful to begin extensive work already by the second part of the first year. There were several reasons for anticipating, in part, this work:

- each group had previous experience with this task, and therefore some preliminary results, having been achieved using rapidly constructed tools for a 'gross' extraction of superordinates, could form the basis for a refinement of the methodology;
- the analysis of the taxonomies obtained from the MRDs is an essential preliminary step for the choice of the vocabulary subset and for further decisions on the analysis of the 'differentia' part of the definitions.

The extraction of taxonomies has been carried out in all the sites by all the partners involved in this work part, though each partner used a partially different work strategy.

Taxonomies have been extracted and built extensively for the whole of each available dictionary in Amsterdam and Pisa. Cambridge and Barcelona chose the strategy of building taxonomy top-down, starting from the most generic genuses of a semantic subset, and therefore taxonomies wre built for the lexical subsets analyzed within the Project. All the sites began with noun taxonomies, though some dictionaries have been completely processed (Van Dale and LDOCE in Amsterdam and DMI and Garzanti in Pisa) for all the major POSs, with automatic procedures. The taxonomies analyzed

within the Project have also been disambiguated, partly by procedures, partly interactively.
In Amsterdam a tool has been developed, running on LDOCE and Van Dale, to browse through stored taxonomies and to compare them cross-linguistically.

The taxonomies extracted are stored in the LDBs, and are therefore accessible for interactive querying, and the ones corresponding to the agreed lexical subset are represented in the LKB, giving rise to the hierarchical structures which allow inheritance.

In the second year and a half, work continued on this WP with the more difficult task of extraction of other types of semantic information from the *differentia* part of the definitions. This task was carried out in Barcelona, Pisa and Amsterdam (both for English and Dutch).

The information extracted had to be modelled in a common Type System, in whose design the three partners above and Cambridge have contributed.

The task of arriving at the definition of this common "metalexicon" where the same features (attributes and range of values) are used for all the languages and all the dictionaries, was by no means an easy task, requiring frequent redesign of previous versions, and consequent recoding of the lexical entries.

We are aware that the Type System, as it exists at present, is a compromise between different exigencies: formalizing as much information as possible, formal consistency, clear semantics of what is represented, usefulness for NLP purposes, theoretical adequacy, etc.

The need to satisfy all these requirements has perhaps constrained the result to a rather impoverished system, which is to be taken as a preliminary hypothesis to be revised and improved upon in the follow-up of the Project.

The relevance of the result is in the fact that a common methodological approach to the lexicon may actually lead to a common representational framework: this achievement is a sine qua non condition for future very large-scale European lexicon building projects.

Variation in the Workplan

We must point out that taxonomy extraction, in Amsterdam and Pisa, has been carried out extensively throughout each dictionary, and was not at all limited to the subset as was written originally in the workplan.

Moreover, this task was partially anticipated, so that we could begin the evaluation and analysis of taxonomies for the choice of the subset (both the criteria and the actual words).

Further work for comparing taxonomies, for merging data coming from different sources (both monolingually and multilingually), for developing strategies for the analysis of the 'differentia', and for discussing the organization of the top and middle level of the taxomonies (both monolingually and interlingually), was carried out on the 'Food & Drink' subset for all the languages involved, and only partially for two subsets of Verbs and for other subsets of Nouns.

WP9. EVALUATION OF LDB/LKB SYSTEMS

Goal

This WP aims at producing a final internal evaluation of the LDB and LKB systems, where problems concerning the representation formalism, the software environment, and the particular Type System designed, are spelled out with respect to the exigencies that cropped up in each site during the completion of the work on the respective lexica and with the different approaches.

In particular it aims at an evaluation of:

- the utility of lexicons for the multilingual testbed derived from the LDB;
- the performance of the LDB/LKB software;
- the quality and utility of the LRL.

Deliverables

The TA predicted a preliminary report at 24 months and a final one at 30 months. Given the agreement with the EC of not having the 2nd year Review, the preliminary report has been incorporated in the final one, thus producing a single deliverable with input from all the partners and edited by Barcelona.
Amsterdam, Barcelona, Cambridge, Dublin, Pisa, "Final Evaluation of LDB/LKB System", Barcelona, May 1992 (Deliverable No. 11).

Responsibility

Barcelona is responsible for reporting on this work part, with the contribution of all the partners.

Status with the reference to the TA

From the point of view of the software tools, the evaluation task has been carried out by every partner all along the project development (in fact, since the 1st release of LDB was provided).

Most suggestions have been incorporated into successive releases of the LDB and LKB software.

As regards the exploitation of the LKB, two different viewpoints have been taken into account:

a) extracting information from LDBs and converting it into the LKB representation system,

b) building lexicons for NLP tasks, extracted from this data.

Dublin has evaluated this 2nd view, whereas the other partners have dealt with the 1st one.

Barcelona has integrated these partial evaluations and produced the report that constitutes the Deliverable No.11.

WP10. REPORT ON FEASIBILITY OF MRD SOURCES FOR NLP SYSTEMS

Goal

The goal is to produce a global internal overview and evaluation of the Project objectives, approach and achievements.

Deliverables

A report was foreseen as a 30 month deliverable.
W. Meijs, "Feasibility of MRD Sources for NLP Systems", Amsterdam, May 1992 (Deliverable No. 12).

Responsibility

Amsterdam is responsible for this deliverable, with the cooperation of all the partners.

Status with reference to the TA

This WP assesses the results of the Acquilex project in the light of the objectives set out originally in the Technical Annexe (TA). The main issue is an evaluation of the Acquilex project in terms of efficiency, cost-effectiveness and benefits for the NLP community.

The second part of the Report, in fact, deals with the central issue: the feasibility, the utility, and the cost-effectiveness of the semi-automatic re-use of MRD data for NLP purposes. It assesses the impact of Acquilex, through its methodology and software output as well as through the many publications on theoretical issues that have come out of it, on the state of the art in NLP-oriented research and applications.

Looking ahead at the future, the results achieved are linked up in section 5 with the plans for Acquilex II, and likely developments in the field of computational lexicology and lexicography to which it will contribute.

B. Concluding remarks

It must be pointed out that this project has been the first initiative where (computational) linguistics has attempted to make a substantial contribution to the study of realistic lexical data on a large scale, in an effort to fill in the wide gap created in the last decades between theoretical linguistics and "real life" of the language, as manifested in the real needs of the actual ordinary language users and workers (among which there are lexicographers). In this respect it has been the precursor of a series of other initiatives where efforts are being made in this direction. We must mention here EUROTRA-7, MULTILEX, GENELEX, some of the ET-10 and LRE projects.

ACQUILEX was also the first large project aiming specifically at developing adequate methodological tools for dealing with the lexicon, in the conviction that a large-scale and systematic linguistic analysis of the lexical system is possible (not limited to a small number of examples) and that this can be based on traditional lexicographic definitions.

Obviously this last assumption has proved only partially correct. Leaving aside well-known problems of inconsistency and incoherency, which are also due to the fact that printed dictionaries are destined for human (not machine) consumption, there is an obvious problem of "incompleteness", in the sense of inadequacy of traditional lexicographic definitions with respect to "all" the requirements of an NLP lexicon. This inadequacy is, however, partial, i.e. even though meanings are certainly not stated adequately and spelled out in all details in any existing dictionary, they are described and characterized in their main features, and this occurs extensively in the entire lexicon. It is this very large bulk of semantic information which has proved to be extractable (by semi-automatic means) and representable in a uniform manner across different dictionary sources and - most importantly - across different languages.

A warning is due here, i.e. these vast amounts of semantic properties and relations become really reusable only if there exists a methodological framework where these data can be meaningfully inserted. This is the real value of the process of integration between computational linguistics and lexicography, a process which began within ACQUILEX and EUROTRA-7, and will receive much more effort in ACQUILEX-II and in other EC funded Lexical Projects which are about to start. A big effort towards this integration is a must if we hope to have within the short- or medium-term large Computational Lexicons really usable for a number of NLP applications.

With respect to the lack of completeness, once it is accepted as a fact and it is recognized as a positive value the consideration about the quantity of semantic data, one can think of which other sources can be used to extract part of what is missing in MRDs. Two obvious sources of complementary lexical knowledge are theoretical linguists' competence on the one hand, and large textual corpora on the other hand. This direction of integration of lexical knowledge coming from different sources, whose necessity became apparent in ACQUILEX, will be fully incorporated in ACQUILEX-II.

2. UPDATE ON THE STATE-OF-THE-ART

The general context

It is the shared opinion of all the partners that no substantial world-wide developments have taken place during the project, which should justify a modification in the direction of our work at the scientific and technical level.
In general, research activities continued along the directions already described in the Technical Annex, and therefore no adjustments seemed to be dictated during the course of the Action.

On recent occasions (COLING, Berlin Conference on MT), a general interest has been declared and specific initiatives have been suggested in the field of the evaluation of NLP methods and systems. Unfortunately, these studies have not yet produced any method or result which can be of help in the strategies of evaluation of the LKB. We hope that this may happen with efforts made by the EC in this area.

Of course, frequent exchanges have occurred with researchers working on related topics outside the project.

New relevant bibliographical issues are included in the various working papers and deliverables annexed to this report.

The European framework: ACQUILEX vs. other Projects

It is instead worth mentioning here some of the developments in the field, at the level of research policy and organisation in Europe and outside Europe, which has modified the general context of R&D in which the project operates, improving the possibilities of exchanges with other ESPRIT and external activities, and greatly increasing the potential value and the immediate reusability of ACQUILEX results. Some of those other projects have explicitly decided to import our results.

Historical background

The Grosseto Workshop "On automating the lexicon", organized in 1986 in the framework of the 3th Multilingual Action Plan, succeeded in putting together, for the first time, linguistics, natural language processing, computational linguistics, artificial intelligence, psycholinguistics, software and hardware companies, publishing houses, etc. One of the results of the workshop was the setting-up of a number of recommendations, which explicitly stated the need for reusable machine readable linguistic resources, and in particular of:

- large lexical databases;
- large textual corpora;
- exchange standards;
- an organizational framework for the cooperation of various categories of producers and users;
- computational tools for the management, access, and processing of linguistic data.

In particular, two complementary aspects of the notion of "reusability" were identified, aimed at:

1) reusing already existing linguistic resources (e.g. MR Dictionaries, textual corpora, thesauri, etc.), possibly in a context different from their original purpose and mainly by extracting or making explicit the information which was only implicit in the original data;

2) building linguistic/lexical resources which can be reused both in a number of different theoretical frameworks and within different applications (not only by procedures but also by human users), i.e. a sort of 'polytheoretical', 'multifunctional' linguistic resource.

The recommendations produced at Grosseto and in various successive occasions constitute, altogether, a general strategical work plan. A series of initiatives, efforts, and proposals aimed at implementing this work plan gave rise to a number of national and international projects.

Overview of Lexical Projects and Initiatives

ACQUILEX has the purpose of exploring, testing and defining the reusability notion in its first sense by reusing data found in traditional dictionaries, for NLP systems.

The ongoing Text Encoding Initiative (TEI), promoted by ACL, ACH, ALLC, supported by the EEC, NEH, Mellon Foundation, and sponsored by 30 major scientific and professional associations (LSA, CLA, APA, EEEI, etc), has the purpose of defining standards for exchanging texts, machine-readable dictionaries included, and the linguistic analyses which are annotated in texts.

The ongoing Survey of Linguistics Resources in MRF, supported by the EEC and promoted by ACH, ALLC, ACL, EURALEX, ESF, NEH, etc., has the purpose of inventoring and describing textual corpora, MRDs, LDBs, terminological collections, and speech corpora, in MRF, as a basis for coordination and standardization.

EUROTRA-7, promoted by the EEC and begun in 1990, was a feasibility study focusing in particular on the second aspect of 'reusability'. In this respect, it aimed at evaluating the feasibility of constructing standardized reusable lexical resources.
It has produced, in cooperation with the survey above, a description of existing lexical resources considered from the point of view of applications, contents, standards, implementation techniques, and organisations.
In its second phase it dealt with the problems and possibilities of standardization of the monolingual and multilingual description of lexical units at the different linguistic levels, and of possible architectures for a reusable resource.

MULTILEX, an ESPRIT project beginning November 1990, aims at:
- defining the model of a LDB;
- exploring lexical sources that are reusable in the implementation of the LDB;
- creating the necessary software tools for the management, access, and processing of the LDB;
- implementing a prototype multilingual LDB, according to a defined standard, for multiple uses.

GENELEX, an EUREKA project involving research and industrial partners from France, Italy, Spain, has the following goals:
- to draw a state-of-the-art in the field of Lexical Resources;
- to define a GENELEX model of a LKB common for all the partners;
- to create tools for extraction of knowledge from textual corpora to enrich the LKB;
- to implement a common lexicon of about 3,000 words;
- to implement extended dictionaries in the 3 languages, which will remain the property of the specific partners who developed them.

Contacts and agreements have been made with the above projects, on the following topics:
- computational model of a lexicon (TEI, EUROTRA-7, MULTILEX, GENELEX);
- procedures for converting MRDs (MULTILEX, GENELEX);
- extraction of semantic information from definitions (MULTILEX, GENELEX);
- LDB/LKB architecture (EUROTRA-7, GENELEX).

ACQUILEX being the first large European project in the lexical area, we must stress that the exchange of concrete results has usually flowed in only one direction, i.e. from ACQUILEX towards the above mentioned Projects. Pisa has been the point of contact through which these exchanges have taken place.

Another project is starting now (an ET-10 project) whose objective is very similar to ACQUILEX, but it is geared to analyze a very peculiar dictionary, i.e. the COBUILD. It will be interesting to assess the possibility of exploiting methods and techniques developed within ACQUILEX and to evaluate their reusability when applied to a different type of dictionary.

Two direct consequences of ACQUILEX have to be mentioned in closing.

In the framework of the ESPRIT-DARPA/NSF cooperation two contracts have been signed (one between Pisa and the EC, one between the Consortium of Lexical Research in Las Cruces and NSF) with the explicit purpose of organizing two workshops, one in Las Cruces in autumn 92 and one in Pisa in 93. These have the aim of promoting the exchange of results between the two groups (a European and an American one) already operating within a similar theoretical framework.

The proposal for a follow-up of the project, i.e. ACQUILEX-II, has been accepted by the EC, and the new project should start soon. This should be taken as proof of the importance of this line of research for future developments in the NLP area.

3. WORK PART SUMMARIES

INTRODUCTION

This section is structured into the following 10 parts.
(The Partner reporting on each part is given in parenthesis)

1 Computational Model of a Dictionary (Pisa)
2 Report on the Lexical Requirements of NLP Systems (Dublin)
3 Lexical Data Base Software and Documentation (Cambridge)
4 Conversion of Machine Readable Dictionaries (Amsterdam)
5 Initial Definition of Vocabulary Subset and Lexical Information for Further Work (Amsterdam)
6 The Testbed NLP System (Dublin)
7 Lexical Knowledge Base System (Cambridge)
8 Semantic Taxonomies (Pisa)
9 Evaluation of LDB/LKB Systems (Barcelona)
10 Feasibility of MRD Sources for NLP Systems (Amsterdam)

For each part we provide a Scientific Work Part Summary - an abstract describing the achievements - produced by the designated partner.

3.1 Computational Model of a Dictionary

3.1.1 Abstract

Aim of the work part

This work part is aimed at producing a definition of the computational model of a dictionary, specifying a general representation for a lexical entry capable of handling a variety of different MRD sources and derived lexicons.

The need to exchange or to share data and results among the partners, and even more importantly the need of storing all the different MRD sources into an identical LDB system, require a common method of representing and interpreting MRD data.

The same, or similar, need is felt nowadays to be an urgent requirement within the research community dealing with texts. The dictionary-handling community is an important subset (as shown e.g. by statistics of OTA) and, in this respect, we must mention the contacts, the interaction, and the exchange of information with the Text Encoding Initiative (TEI) and with EUROTRA-7.

Deliverable

This work has resulted in a report which constitutes a 6 month Deliverable:
N. Calzolari, C. Peters, A. Roventini, Computational Model of the Dictionary Entry, Preliminary Report, Pisa 1990.
An evaluation of the model was carried out at the end of the project, whose result is given in a brief report containing the Final Proposal.

Steps of the work part

In order to achieve the result of defining a common method of representation and interpretation of MRD data, a detailed analysis of the structure and content of all the project dictionaries has been carried out. This part of the work was developed within each site for the 'local' MRDs.

During a project meeting in Pisa, a proposal was made concerning the method of representing MRD data in a uniform way. The proposed model was discussed by all the partners (via e-mail and in a meeting in Amsterdam), revisions and modifications were agreed on (also to accomodate features characteristic of each dictionary), and in its present form the model has been used for a uniform description and representation of all the MRD sources available to the project.

At the Amsterdam meeting we also decided jointly on the preliminary format for the Common Lexical Entry (CLE), to which all the MRDs should be converted in, particular for the purposes of building the common Lexical Knowledge Base (LKB) and for the NLP tasks of the testbed system. Especially for this application, it was judged useful to have word-sense based entries, with most of the information based on the word-sense level.

This CLE model has been implemented for all the dictionaries to deal with the tasks for which it was conceived.

Usability also outside of the project

We did not judge it necessary for this project to develop a common 'formal' representation language for MRD data since it is not necessary to exchange MRDs themselves, but only the already 'analysed' MRD data. We therefore need more a model and a method for a uniform interpretation of analysed data rather than a formalism for a uniform encoding of the source data. We thus limited ourselves to the definition of a model of an 'explicit' method for representing dictionary data.

We think that this model can be used as a starting point for a fairly straightforward translation, e.g. in SGML, which, according to us, should be more in accordance with the task of the TEI. This model has been in fact used by the Printed Dictionary Work Group of the TEI, as one of the bases of the TEI proposal for encoding MRDs.

This representation language has been used and/or evaluated by other research groups working on MRDs (e.g. in Prague, Stuttgart University) and is judged to be both easily mappable and compatible with other models (e.g. the IBM Hawthorne).

It has been incorporated, as a proposal to be considered in the feasibility study for reusing lexical resources, in the Survey part of EUROTRA-7.

Summary of the deliverable content

The description of the computational model of the dictionary entry consists of two separate sections.

The first part, Section 1, presents a method which can be used to represent, in a uniform way, the content and structure of the entries of machine-readable dictionaries (MRDs), and contains an explicit standardized representation of the content of the different dictionaries being used within the Project. This representation model can be proposed:

a) as a common basis to be integrated with other specific observations on other dictionaries, or dictionary types, in order to become a general representation model;
b) as a basis which, being explicit enough, can be easily encoded in SGML tags in order to have a TEI-conformant document.

The second part, Section 2, consists of a description of the Common Project Lexical Database Entry. This description is more complex than that of Section 1 because it contains not only a description of the Entry following the same formalism used in Section 1, but also a brief specification of the database model which has been decided on by the Project. Even though the Common Lexical Entry was intended only to be adopted internally to the Project, it is easy, if desired, to transform any dictionary, whose content is represented according to the model presented in the 1st Section, into this CLE.

The two sections together constitute the integrated proposal for the Computational Model of a Dictionary Entry.

3.2 Lexical Requirements for NLP systems

3.2.1 Abstract

This task began with an evaluation of the lexical requirements of NLP. The evaluation then proceeded by dismissing the use of an approach which considers requirements on an application-by-application basis on the grounds that there would be a great deal of unprofitable repetition.

Instead, two sets of NLP tasks were identified: one set of relatively superficial tasks, which are for the most part undemanding in terms of lexical information required, and another set of eight knowledge-based tasks.

Of the second set, two tasks in particular were found to require a great deal of lexical information: parsing of a single sentence, in which we include meaning analysis; and generation of single sentences. The lexical requirements of these tasks are discussed at length in Deliverable No.2.

The other knowledge-based tasks, with the exception of language acquisition, presuppose that single sentence parsing and/or generation tasks have already been carried out. For the most part, they operate on the results of parsing, and so are to a degree insulated from the lexicon. Nevertheless, it was also discovered that they require additional specific kinds of information, some of which is strictly lexical and some of which is more properly thought of as tied to representation-language concepts rather than directly to lexical entries.

Some kinds of information were identified which we do not expect to be derivable from MRDs, at least not in a form which is sufficiently explicit to be deployed in NLP systems. One such example is the semantic effect of derivational affixes; another is general real-world knowledge.

We also reported on the number of ways that NLP systems might wish to access a lexicon, which have an impact on the ways that entries in a lexical knowledge base should be indexed and organised.

3.3 Lexical Data Base Software

3.3.1 Abstract

Introduction

The Lexical Database System (LDB) is a computer system (described fully in the documentation for the LDB, Deliverable No.4) which provides flexible access to machine- readable dictionaries. It supports a user in formulating queries to retrieve subsets of entries from one or more dictionaries, implements the efficient retrieval of entries (using multiple-level indices and ideas drawn from current database technology), and allows new "derived" dictionaries to be created containing entries from a source dictionary, augmented and enriched with new information.

In "Database Models for Computational Linguistics" (*Proceedings of EURALEX-90*, Malaga, Spain), B. Boguraev, E. Briscoe, J. Carroll and A. Copestake identify four classes of dictionary models. The first of these follows the well-established notion of relational databases, mapping dictionary entries into a set of tables. Although this relational model of the lexicon can take advantage of established database technology, it is generally agreed to be unsuitable for mapping dictionaries into, given the intricate nature of and subtle interactions within, lexical data. The second class is the hierarchical model which employs a structured representation to encode the complex structural relationships between the fields of entries (exploiting the insight that dictionary entries can naturally be regarded as shallow hierarchies with an indefinite number of attributes at each level). The third class is the tagged model; in contrast to the hierarchical model which fails to preserve the visual, human-readable interrelationships amongst the contents of dictionary entries, this model places emphasis on preserving all of the information associated with the original printed form of the dictionary entry, but in the process fails to offer a natural way of making explicit statements concerning the implicit structural relationships of the elements within the entry.

The fourth class of dictionary model, the two-level model, aims at combining the advantages of the hierarchical and tagged models. This is the model implemented in the LDB. 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. In the LDB, the "mounting" of a new machine-readable dictionary consists mainly of defining what these indices are, how they are to be extracted from entries, and then telling the system to create permanent files on disc holding the indices. 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.

Integrated into the LDB is the Flexible Pattern Matching Parsing Tool (FPar), a parser designed for the specialized purpose of parsing dictionary entry definition texts.

Querying a Dictionary

A 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'. A secondary task in mounting a dictionary is to define 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.

Once a dictionary has been mounted (its index files created and format for queries defined), it may be 'loaded' into an LDB session. Loading a dictionary causes its index files to be located and prepared for use, and menus of possible values for query attributes to be created. Once a dictionary has been loaded, the user may interactively construct one or more queries relating to that dictionary, and then for each query ask the system to retrieve all the entries which satisfy it. (The next section outlines the extensive facilities provided to the user by the LDB for constructing and modifying queries). Several dictionaries can be loaded in the same session: they are all available for access concurrently.

Looking a query up is a two-stage process. The LDB first maps the query into a collection of indices, determines which of these are the most discriminating (i.e. have the lowest frequency, based on statistics which were gathered during the creation of the index files), and finds an initial set of entries which satisfy this subset of indices by computing the intersection of the pointers from the index files to entries in the dictionary corresponding to the indices. The LDB then retrieves this set of entries from the source dictionary, checks which ones satisfy the rest of the (less discriminating) indices, and returns the ones which do as the final result. Crucial to the efficiency of dictionary query lookup is a good partition of the indices in the query into those used to form the initial candidate set of entries and those used to check these entries after they have been retrieved. The LDB bases its partition on estimates of the relative costs of reading and intersecting entry pointers versus reading and checking the entries themselves, on the numbers of pointers that will be read, and the expected probability of an entry succeeding in a check against a particular index.

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). The LDB offers a number of options for the display of answers to a query: after informing the user of how many entries and senses satisfied the query, the LDB can be asked to display the headwords of all the results, of a sample of them, of the first one, or nothing. In addition, a user-defined option allows any portion of result entries to be displayed rather than simply the headword, and results can also be returned as entry pointers to allow arbitrary set operations (e.g. union, intersection, set difference) to be applied to them.

Derived Dictionaries

As well as supporting several dictionaries being available for access concurrently, the LDB allows the user to apply a single query to two or more dictionaries simultaneously (as long as all of the dictionaries concerned are derived from a single 'source' dictionary). In fact, one of the dictionaries may be the source itself. A derived dictionary will typically consist of an elaboration of a subset of the information in the source dictionary: for example containing just the definition part of entries in the source, but having parsed representations of the definitions. The LDB also makes it straightforward to create derived dictionaries based on the entries returned from the lookup of a query.

Formulating Queries

As mentioned above, a query is represented as a hierarchical collection of attributes with associated values. The basic values for attributes are usually atomic tokens (e.g. numbers, dictionary codes, or words) as in the example above. More complex types of value may be made out of these basic values, however, in order to express conjunctive and disjunctive queries and (atomic) negation. Basic values can be wildcarded with '?', matching any single sub-element, and '*' matching any sequence of sub-elements.

In addition to the attributes defined during the mounting of a dictionary, the LDB itself provides ones called 'headword' and 'constr' (short for constraint). 'Headword' allows a (partial, using wildcards) specification of the headword to be made on entries that will be retrieved. 'Constr' provides a way of expressing queries which cannot be formulated in a simple attribute-value form. The attribute can have one or more (conjunctive) values, each either a disjunction of a set of, or negation of a, dictionary index specification or call to Lisp.

The LDB gives the user the option of using either a TTY or a graphical interface for constructing queries, modifying them, looking them up, reading them from a file, and saving them back to the disc. Both interfaces provide full facilities for the quick and accurate manipulation of queries. The LDB graphical interface on the Apple Macintosh makes extensive use of the mouse, windows, and pop-up and pull-down menus.

Status of the LDB Deliverables

In accordance with the Technical Annexe, the Lexical Database System and Flexible Pattern Matching Parsing Tool have been distributed to all collaborating ACQUILEX project sites in the form of executable software and full documentation; in April 1990 as Deliverable No.3 and in October 1990 as Deliverable No.4.

The Deliverable 4 technical documentation, consists of two parts:

a) the report of J. Carrol, "Lexical Data Base System User Manual", concerning the Cambridge LDB system, to which the preceeding description refers in particular. This system is the common project software being used by all the partners;

b) the report of E. Marinai, C. Peters, E. Picchi, "Pisa Multi-Lexical Databases: an integrated system for the acquisition, maintenance and interrogation of mono- and

bilingual LDBS". The Pisa system offers similar functionalities to the Cambridge system and, according to the technical annex, has been developed under MS\DOS.

Existing users of the LDB include Dr. Beth Levin's dictionary research group at North-Western University, USA, Professor William Marslen-Wilson's psycholinguistic research group at Birkbeck College, University of London, and Branimir Boguraev of the Lexical System Group, IBM T. J. Watson Research Center, New York, USA.

3.4 Conversion of Machine Readable Dictionaries

3.4.1 Abstract

By and large, the goals of this Work Part have been achieved for the 12 month stage according to plan along the lines suggested in the Technical Annexe. Two different approaches were discussed to coordinate the conversion:

1) developing a common general parsing tool which could then be used by all sites;
2) designing and writing specific software tools for parsing and converting each dictionary.

Because of the restricted time schedule, work on the conversion of individual MRDs had to start right away, necessarily using local software and methods. From these locally-handled conversions it became clear that the raw form of the various MRDs differed widely (varying from virtual typesetting tapes to almost database format). In this respect developing general software, although an intellectual challenge, would be too time consuming and would have lead to a serious delay in the availability of the data.

The sites which brought MRDs into the project all managed to convert their MRDs and load the results in the LDB. This in spite of the fact that there were very substantial differences both in the difficulties to be overcome (due to widely varying levels of structural coherence and consistency in the source-materials), and in the varying computational methods and techniques adopted in the different sites (due to the diverse computational experiences, resources and philosophies in the different sites).

Initially, the MRDs available at the various sites were pre- processed using existing local software. This initial processing mainly involved nitty-gritty groundwork: 'cleaning up', error- correction, systematization, indexing, 'Lispifying', consistency- checks etc. Meanwhile the joint work on Deliverable 1 had led to a first definition of the CLE, and the first version of the LDB software developed by Cambridge was made available, followed by a series of updated versions.

In practice the conversion was very much an interactive and partly cyclical process: a first attempt to load a particular MRD into the LDB would fail, or give unsatisfactory results on specific points. This would then lead to suggestions for adaptations in the LDB package and/or changes in the data-type distinctions in the local MRD, or in its mapping onto the CLE, followed by a new loading operation which would generally lead to improved results. Remaining problems would be tackled in a further cycle and thus the conversion process would proceed in a similar trial-and-error manner until the results were deemed satisfactory. The advantage of this approach was that it forced participants to really 'come to grips' both with the general structure and with the finer details of their MRDs, while at the same time leading to significant improvements in the LDB software. The conversion process as such thus turned out to be a significant developmental and 'learning' stage in the Project, and the results (i.e. optimally accessible data across different MRDs) functioned as reliable sources for the further type of processing foreseen in the rest of the Work Plan.

As a result of the conversions the following lexical resources are currently accessible in LDB-shells:

English monolingual:

- The Longman Dictionary of Contemporary English (LDOCE)
- The Longman Lexicon
- The Oxford Advanced Learners Dictionary (OALD)

Spanish monolingual:

- Vox: Diccionario Ilustrado de la lengua Espaola

Italian monolingual and bilingual:

- Italian Machine Dictionary (IMD)
- Garzanti Nuovo Dizionario Italiano
- Collins Concise Italian/English, English/Italian Dictionary

Dutch monolingual and bilingual:

- Van Dale Groot Woordenboek Hedendaags Nederlands (VDL NN)
- Van Dale Groot Woordenboek Nederlands-Engels (VDL NE)

Together they constitute more than half a million dictionary entries distributed over four languages for which the comparison, extraction and transfer of data can be performed. As such, this collection is probably the most extended resource of lexical knowledge developed world-wide so far.

3.5 Initial Definition of Vocabulary Subset

3.5.1 Abstract

The purpose of selecting a testset was to investigate, within the limits of time and resources set for the project, the feasibility of extracting lexical knowledge from MRDs which could be of help in NLP and the building of a multilingual LKB for a common subset of the vocabularies.

The definition of a sensible testset of, on the one hand, a sufficient number of vocabulary items (or senses, rather - with a mark set, rather arbitrarily, at 2000), and, on the other hand, of an appropriate range of grammatical characteristics met with considerable difficulties at first, mainly due to the fact that the timing of this deliverable within the overall workplan was somewhat ill-conceived. Basically what was involved here was a typical "bootstrap" problem: deciding on an interesting vocabulary subset wasn't really possible before the successful extraction of reliable and sizeable taxonomies, and taxonomy-extraction had been planned after the selection of the vocabulary-subset. Similarly, while a report on lexical requirements in NLP systems stating in general terms what kinds of grammatical information might be needed in NLP systems had been produced as one of the earliest deliverables (Del. no. 2), the specific grammatical aspects that could reasonably be dealt with within the Acquilex NLP testbed could only be determined once the vocabulary testset had been decided on.

In the course of the project these problems evened out, though, and after the taxonomy hurdle had been taken, sufficiently large and interesting vocabulary subsets and associated grammatical characteristics could be selected. The choice of the specific subsets that were singled out (food and drinks, movement verbs etc.) was motivated on the one hand by our assessment of sensible NLP requirements, on the other hand by the consideration that these sets were likely to bring up specific interesting aspects that would have direct relevance also for current theoretical issues.

This expectation was brought out, and has led to a large number of publications on such issues as metaphoric and metonymic sense-extensions, lexical rules like portioning and grinding, the systematics of movement and psych verbs etc. In the end, then, we think we managed to arrive at a reasonable balance between the two seemingly conflicting aims of representativeness and comprehensiveness that we had set ourselves in the initial subset definition.

3.6 NLP Testbed system

3.6.1 Abstract

The research reported in the deliverable "Lexical requirements of Natural Language Processing" aimed to establish the range of types of arguably lexical information that would be useful to practical NLP systems. It considered a full range of NLP tasks. Two of these tasks, sentence analysis and sentence generation, are clearly and intimately dependent on lexical information; but other tasks, often involving discourse-level phenomena, also depend on information that is arguably lexical in nature. The workpart goals included assessing the utility for NLP of the information extracted from dictionaries and represented in the LKB. This work contributed to achievement of those goals in that it established ideal criteria for such an assessment.

The work on the testbed machine-translation system can be broken into three phases. The first phase was the adaptation of an existing English sentence analyser, in order to bring the structure of its lexicon more into line with the structure envisaged by the "Computational model of the lexicon", but still using hand-coded lexical information. The second phase was the construction of two sentence generators, for Dutch and for Italian, to be compatible with the English analyser, again using hand-coded lexical information. These combination of the products of these two phases resulted in the prototype machine translation. All this work was however merely preparatory to the third phase, the exploitation of lexical information that had been extracted from dictionaries.

This final phase involves interfacing the LKB (lexical knowledge base) to the testbed system, in order that hand-coded information may be supplanted by dictionary-derived information. The major goal of the workpart is the assessment of the utility of this information for practical NLP. The work therefore leads directly to the achievement of this goal, since the tasks involved in translating from one language to another constitute a very large portion of the overall NLP range. It will allow assessment of the utility of the lexical information according to the following criteria:

- The extent to which dictionary-derived information can supplant hand-coded information
- The ease with which dictionary-derived information can be extracted from the LKB and deployed in another system; and hence whether the goal of reusability of lexical information resources has been met
- The quality of the dictionary-derived information as compared with what is realistically achievable in hand-coding

3.7 Lexical Knowledge Base

3.7.1 Abstract

This WP refers to work on the LKB software system, the conversion tools, and on the representation of information in the LKB.

The LKB system has been designed to allow the representation of syntactic and semantic information extracted from MRDs. The LKB supports the representation of reusable multilingual information in a structured lexical knowledge base. Substantial monolingual lexicon fragments have been derived (semi-)automatically for English, Spanish, Italian and Dutch and represented in the LKB, and cross-linguistic links between these entries have been generated semi-automatically. Working paper 036 (Copestake, 1992a) is a relatively general paper describing the LKB software and aspects of its use.

LKB software

The basic LKB system implements a general-purpose lexical representation language (LRL) which supports the general development of lexicons using typed feature structures, default inheritance, lexical rules and translation links. The LKB has a menu-driven, graphical user interface, and uses a variety of techniques to aid the linguist in developing large scale lexicons.

The initial version of the LKB system was released at the 18 month point. An international workshop on default inheritance in the lexicon was held in Cambridge to coincide with the release. The proceedings of this are to be published by CUP. The second release of the LKB system took place in August 1991; this incorporated tools for multilingual work. Minor updates to the system have been released since then. Documentation for the various releases has also been issued, as has an implementation outline for those who wish to interface software to the LKB. A version of the system, including illustrative type systems, lexicons etc., was demonstrated at ANLP-92 and a stand-alone, demonstration system is generally available and has been distributed to other researchers.

Representation in the LKB

Representation of information derived from MRDs in the LKB depends on the prior development of an appropriate theory and its description in the typed feature structure framework.

LDB to LKB conversion

The full LKB is integrated with the wide range of other software which has been developed on the ACQUILEX project. A range of tools have been developed to allow information stored in the LDB to be converted into LKB entries. The Dictionary Correlation Kit correlates (related) MRDs semi-automatica lly. This has been used to augment the Longman Dictionary of Contemporary English (LDOCE) with semantic information derived from the Longman Lexicon of Contemporary English, allowing LKB verb entries containing detailed semantic and syntactic information to be derived. This approach does not rely on the use of taxonomic information, which has been found to be less useful for verbs than it is for nouns.

We have investigated a variety of approaches to extracting information from the dictionary definitions, which do make use of taxonomies built from sense-disambiguated genus terms, and of information extracted from the differentia. The flexible pattern matching / parsing tool (FPar) which is integrated with the LDB is based on the system described in Alshawi (1989). The SEISD system developed in Barcelona integrates FPar with a morphological analyser (SEGWORD) and a taxonomy building program in order to produce LKB representations from the Spanish Vox dictionary.

General purpose parsers with special purpose grammars have also been used to analyse definitions. Conversion programs parametrised according to the particular semantic subset of definitions being investigated allow LKB entries to be produced from the output of these parsers.

The multilingual LKB

The work referred to in the section above has resulted in the construction of substantial monolingual lexicon fragments in the LKB in all four languages. However the LKB allows the representation of complex cross-linguistic relationships. An interface to bilingual dictionaries has been combined with a general feature structure matching utility, LUCIFER, to allow the semi-automatic generation of translation links between word senses in the monolingual LKB lexicons in the four project languages. Much recent work has been devoted to experimenting with the representation language and the mapping software in order to generate a multilingual, linked, LKB. The results of this have not yet been reported in detail in papers.

Use of bilinguals

The English-Dutch and Dutch-English Van Dale bilinguals were mounted in the LDB in order to test the mapping software. In Barcelona no machine readable bilingual was available in time, and thus bilingual links between word forms were extracted by hand for the subset of the vocabulary with which we are concerned.

Tlink creation software

The original program for creating tlinks adopts a statistical approach to the problem of sense disambiguation. The original tlink program essentially created simple-tlinks by consulting the bilingual dictionary, and the source and target monolingual LKB lexicons. This software has been extensively augmented by the Barcelona team in order to extend the coverage of automatic tlink creation in order to cope with some more complex tlinks that can be be built up in an automatic fashion. All the modifications are included in the existing semiautomatic process, so that the user's navigation in the interactive windows remains unchanged.

Results

Details of the results of the preliminary experiment on tlink generation are included in the evaluation document, but are summarised here. In all cases the experiments were performed on limited subsets which were chosen to be compatible. This has meant that the statistical aspect of the mapping software has not been thoroughly tested, since the lexicons were in effect preselected to contain only the correct senses in most cases of ambiguity. However the fact that we can perform such selection and would reflect such

subsets in the type system, in itself suggests that the sense disambiguation process would be successful on larger trials.

3.8 Semantic Taxonomies

3.8.1 Abstract

This WP can be subdivided into two major parts, which have been carried out in two subsequent phases, i.e. i) extraction and building of semantic taxonomies, and ii) extraction of other semantic information from the *differentia* part of the definitions, to be represented in the Type System in the LKB in the form of Feature Structures.

Taxonomies

This work part aims at the extraction of semantic taxonomies from all the project monolingual MRDs for the agreed vocabulary subset. These taxonomies constitute the basis for building an initial version of the Lexical Knowledge Base.

The production of semantic taxonomies for the agreed vocabulary subset was foreseen in the general Workplan as an 18 month Deliverable.

The partners involved in this task are Pisa, Cambridge, Amsterdam, and Barcelona, each having the task of extracting taxonomies for the respective dictionaries.

All of the groups have considered it useful to begin working rather extensively on this work part during the second part of the first year. There were several reasons to partly anticipate this work:
- each group had previous experience with this task, and therefore some preliminary results could be achieved without much effort;
- the methodology for a 'gross' extraction of superordinates (without further refinements) is rather simple;
- the analysis of the taxonomies obtained from the MRDs is an essential preliminary step for the choice of the vocabulary subset and for further decisions on the analysis of the 'differentia' part of the definitions.

The extraction of taxonomies has been carrried out in all the sites by all the partners involved in this work part, though with partially different work strategies.

In Pisa: complete and automatic extraction (for all the definitions of the DMI and of the Garzanti), with disambiguation of the superordinates of the selected subset.

In Cambridge: partial extraction, for specific semantic classes, for the LDOCE, and the superordinates have been (partly by procedures, partly interactively) disambiguated.

In Amsterdam: complete extraction for the LDOCE nouns and verbs, with almost complete disambiguation of the nominal superordinates. Adjectival definitions have been analysed as well, but have not been stored in the form of a taxonomy because their definitions do not provide taxonomy information in the same way as noun and verb definitions do. Complete and automatic extraction for the Van Dale nouns, with work on semantic disambiguation for the selected subset.

In Barcelona: extraction and disambiguation of the selected subset from the Vox.

The taxonomies extracted have also been stored in the LDBs, and are therefore accessible for interactive querying. Additionally, Amsterdam has developed a tool to

browse through stored taxonomies derived from LDOCE and Van Dale and to compare them cross-linguistically. The possibility of interactive access to taxonomies in the LDBs has allowed carrying out rather detailed analyses of specific semantic subsets, which were essential both for defining the criteria of choosing the vocabulary subset, and for beginning the design of subsequent work aimed at the construction of the LKB (e.g. taxonomy comparison and/or merging, their reorganization, analysis of the 'differentia', etc.).

We must point out that taxonomy extraction, both in Amsterdam and Pisa, has been carried out extensively throughout each dictionary in its entirety, and was not at all limited to the subset as was written in the workplan and in the description of the corresponding deliverable. This was, in fact, the only methodologically sound way of accomplishing this task (given that the genus term of each individual entry can belong to any part of the lexicon itself, the same being true for its genus term, and so on).

The advantage of having partially anticipated this part of the workplan was that we had material available:
- to be evaluated for the choice of the subset;
- to be compared for merging data coming from different sources, both monolingually and multilingually;
- to be used to develop strategies for the analysis of the 'differentia';
- for discussing the organization of the top level of the taxomonies and possibly of the middle level, both monolingually and interlingually.

Other semantic information from the differentia

The *differentiae* of all the definitions of the FOOD and DRINKS subsets have been analyzed for the four languages: in Amsterdam the LDOCE and Van Dale, in Barcelona the Vox, in Pisa the Garzanti and DMI.

The analysis of the definitions led to the individuation of a set of typical frequently recurring patterns, which can be used to design the basic template for the meaning type FOOD. This template has two main purposes: i) to guide the automatic semantic interpretation of the definitions, and ii) to guide the design of the Type system whose feature structures are the result of the conversion of these basic templates.

A core set of patterns were present in each dictionary, differences being more evident in peripherical types of information.

The automatic analysis of the definitions with the purpose of identifying and extracting the relevant features and values according to the preliminary template defined for each semantic area was carried out in each site in an independent way. This decision was taken because of lack of time and money in imposing the use of a unique methodology. Each site made the choice of the parsing strategy according to local experience and availability of tools.

Barcelona used a syntactic-semantic parser designed by Alshawi, while both Amsterdam and Pisa adopted a two-step procedure (a syntactic parsing followed by a semantic interpreter based on a pattern-matching technique), but they decided to adapt tools independently developed by the respective groups.

In all sites the third step, operating on the output of the syntactic/semantic analysis, is the conversion of extracted information into the common Type system, thus creating the lexical entries for the LKB.

These extraction and conversion phases were rather expensive and time-consuming, one advantage being however that these procedures are now available for being applied to other semantic subsets and other POSs without starting from scratch, but with only the need of revisions, adaptations and refinements to the existing modules.

It is this common set of Types (feature structures made up of the same attributes and values) which constrains the information which is extracted, and forces one to represent the raw textual data in an explicit, consistent, and uniform way, thus aiming towards a standardization.

The hierarchical structure provided by the taxonomic data in the LKB gives the possibility of deriving massive information through inheritance, with very little locally specified information.

3.9 Evaluation of LDB/LKB Systems

3.9.1 Abstract

Main goals of this Workpart deal with the evaluation of LDB/LKB System as regards the accomplishment of the overall objectives of the Project.

LDB/LKB currently includes software and lexical information extracted from Dutch, English, Italian and Spanish dictionaries.

LDB/LKB system involves two core software systems, the nuclear LDB and LKB, and other programs covering different functionalities related to the extraction task, as Taxbuild, SEISD, etc.

Lexical information includes the different LDBs, derived dictionaries, Type Structure Content and the Multilingual LKB.

LDB Software

Two LDB software systems were developed, by Cambridge and Pisa, in the early stages of the Project. A 1st release was provided to all the partners and reported as the 6 month Deliverable No. 3. A 2nd release was delivered and reported as the 12 month Deliverable No. 4.

Every partner has made an extensive use of Cambridge software in order to build its own LDB for representing lexical data extracted from the respective MRDs.

Most suggestions made from users have been incorporated in succesive releases of LDB software. Now, we can consider the LDB software as a very useful and friendly environment for lexicographic and lexicologic work and research.

Complementary LDB tools

Some tools were produced at the early steps of the project and were integrated with the LDB to facilitate acquisition of information from the MRDs previously described. They include:

* FPar (Carroll 1990), pattern-matching based parser used for definition analysis.
* Segword (WP 4), morphological analyser.
* Taxonomy creation tool (WP 12).

Furthermore, new tools were developed to deal with more specific problems or to provide an environment to integrate processes involving both LDB and LKB.

LDB data

The LDB software has shown its robustness allowing the loading and indexing of several very distinct dictionaries.

Different facilities of LDB software have been widely used and appreciated by every site.

An explicit quotation must be made for the construction of derived dictionaries. Amsterdam and Cambridge have used this facility and applied it to the construction of the LKB.

LDB evaluation

In general, the performance of the system has been considered highly satisfactory in all the three modes of interaction:

* graphical access mode
* command mode
* low level software access mode.

Basically the LDB has been used as a stand-alone system, though LDB facilities have been widely used as pieces for higher level software systems (as SEISD and DCK).

LKB software

The distinction between an LDB and an LKB is discussed in Briscoe (WP 41). The LKB and LDB can be combined to facilitate the process of acquiring information from MRDs, even though both can be used as stand-alone systems.

The LDB and the LKB are thus complementary; unlike the LDB, the LKB cannot flexibly support large scale database queries and is not intended to store not analysed (or partially analysed) data.

The initial version of the LKB system was released at the 18 month point. The representation language was described in the document "Functionality of the LKB".

Many efforts have been made by all the partners on establishing the basis of the LKB structure and the form of the LRL. The use of typed feature structures is based on Carpenter (1990).

The second release of the LKB system took place in August 91; it incorporated tools for multilingual work (see WP 43). The LKB's lexical rule mechanism has been described in WP 22; the issues involved in the use of the representation language for semi-automatic acquisition of large-scale data are discussed by Copestake.

LKB complementary tools

Several software tools have been developed around the LKB nucleous. Broad purpose environments, as SEISD, or more specific tools for mapping lexical entries, t-links generation, learning, etc.

Type Structure

Most critics about LKB refer more to the Type System content than to the LKB phylosophy or the underlying LRL.

The methodological approach followed in Acquilex, i.e. incremental development of both type structure and LKB construction, starting with a small subset of concepts (selected taking into account a variety of heterogeneous criteria, see Deliverable 6) has lead to a generally sufficient coverage of the Type System (about 640 Types) but with local limitations (for instance, a lack of adjectives and verbs appearing in noun definitions).

From LDB to LKB

Due to the characteristics of MRDs as source of information, an incremental and semi-automatic approach is needed in order to perform the task of extracting and formalysing the information to build lexicons for NLP.

A variety of techniques, methods and tools (as referred above) have been used in different sites for such process.

LKB evaluation

Two different views have been stated as regards the LKB evaluation. The adequacy of the software tools and the methodology to achieve the multilingual LKB construction and the feasibility of such knowledge structure to obtain computational lexicons to be used in NLP applications.

Dublin has been involved with the last view whereas the other partners have dealt with the former (see Deliverable 11 for a more detailed balance of these views).

A special mention must be addressed to the use of LKB for the semi-automatic generation of translation links (t-links) between different language lexical entries.

Deliverable

This work has resulted in the 30 months deliverable "Final Evaluation of LDB/LKB system" (No. 11).

Future developments (for a follow-up)

LDB is basically a consultation device. It provides flexible access to lexical entries on the basis of queries involving any information contained in any related MRD source. On the other hand, it does not provide facilities for editing or changing the source data, and therefore, the application of LDB for lexicographic building purposes is limited.

On the other side, LKB provides a representation language and defines valid operations on entries. It provides facilities for creating type systems, loading lexicons, displaying fully expanded FSs, type checking, and so forth.

An extension of the functionalities of both LDB and LKB, allowing constructive operations for the former and database-like access for the later would be desirable and have been included in the Acquilex II proposal.

3.10 Feasibility of MRD Sources for NLP Systems

3.10.1 Abstract

The Feasibility Report assesses the results of the Acquilex project in the light of the objectives set out originally in the Technical Annexe (TA). Although it includes a brief review of the main components of the project corresponding to the various Deliverables listed in the TA, this is meant only to provide the background for the main issue of this Report, viz. an evaluation of the Acquilex project in terms of efficiency, cost-effectiveness and benefits for the NLP community.

In the introduction (section 2) the main goals and overall design of the project are outlined. Section 3 discusses the various substantive subgoals of the project corresponding to the specific deliverables listed in the Technical Annexe. Section 4 deals with the central issue: the feasibility, the utility, and the cost-effectiveness of the semi-automatic re-use of MRD data for NLP purposes. It assesses the impact of Acquilex, through its methodology and software output as well as through the many publications on theoretical issues that have come out of it, on the state of the art in NLP-oriented research and applications. Looking ahead at the future, the results achieved are linked up in section 5 with the plans for Acquilex II, and likely developments in the field of computational lexicology and lexicography to which it will contribute. Following the normal Reference section there is a Bibliography listing all the publications that have resulted directly or indirectly from the Acquilex project. The Appendix gives some concrete illustrative facts and figures about the work performed in some of the sites.

4. DELIVERABLE OVERVIEW

Deliverables of the 1st year

1) N. Calzolari, C. Peters, A. Roventini, Computational Model of the Dictionary Entry, Preliminary Report, Pisa, April 1990.

2) A. Cater, C. Guo, Lexical Requirements of Natural Language Processing, Dublin, July 1990.

3) J. Carroll, Lexical Database System User Manual, Cambridge, April 1990.

4) J. Carroll, Lexical Database System User Manual, Cambridge, October 1990. E. Marinai, C. Peters, E. Picchi, The Pisa Multi-Lexical Database System, Pisa, November 1990

5) Conversion of Machine Readable Dictionary Sources, Amsterdam, November 1990.

6) Initial Definition of the Vocabulary Subset, Preliminary Report, Amsterdam, November 1990.

7) A. W.S. Cater, Testbed English Language Analyser, Dublin, November 1990.

Deliverables of the 2nd year and a half (here enclosed)

8) T. Briscoe, A. Copestake, LKB, Cambridge, May 1992.

9) N. Calzolari, T. Marti, P. Vossen, Taxonomies and Feature Structures, Pisa, November 1991.

10) A. Cater, P. Matthews, NLP Testbed and LKB, Dublin, June 1992.

11) Amsterdam, Barcelona, Cambridge, Dublin, Pisa, Final Evaluation of LDB/LKB System, Barcelona, June 1992.

12) W. Meijs, Feasibility of MRD Sources for NLP Systems, Amsterdam, May 1992.

5. WORKING PAPERS

Working papers of the 1st year

Vossen P., October 1989, "Polysemy and vagueness of meaning descriptions in the Longman dictionary of contemporary English", (also published in: J. Svartvik and H. Wekker (eds.), *Topics in English Linguistics*, Mouton de Gruyter, The Hague) ESPRIT, BRA-3030 ACQUILEX WP NO.001

Vossen P., December 1989, "Getting to grips with the structure of the VanDale Dictionary", ESPRIT BRA-3030 ACQUILEX WP NO.002

Condoravdi C., March 1990, "Symmetric Predicates, Verbal Classes & Diathesis Alternations", ESPRIT BRA-3030 ACQUILEX WP NO.003

Sanfilippo A., March 1990, "A morphological Analyser for English & Italian", ESPRIT BRA-3030 ACQUILEX WP NO.004

Calzolari, N. and A. Zampolli, April 1990, "Lexical Databases and Textual Corpora: a trend of convergence between Computational Linguistics and Literary and Linguistic Computing", (also in S.Hockey and N.Ide (eds.), *Proceedings of the ALLC/ACH Conference*, Toronto, Canada), ESPRIT BRA-3030 ACQUILEX WP NO.005

Calzolari, N., May 1990, "Lexical Databases and Textual Corpora: perspectives of integration for a Lexical Knowledge Base", (in U.Zernik (ed.), *Proceedings of a Workshop on Lexical Acquisition*, Detroit, MIT Press), ESPRIT BRA-3030 ACQUILEX WP NO.006

Condoravdi C. and A. Sanfilippo, July 1990, "Notes on Psychological Predicates", ESPRIT BRA-3030 ACQUILEX WP NO.007

Copestake A., July 1990, "An approach to building the hierarchical element of a lexical knowledge base from a MRD", (Paper presented at the International Workshop on Inheritance in Natural Language Processing, Tilburg, 1990), ESPRIT BRA-3030 ACQUILEX WP NO.008

Vossen P., and I. Serail, July 1990, "Word-Devil: A taxonomy-browser for decomposition via the lexicon", ESPRIT BRA-3030 ACQUILEX WP NO.009

Vossen P., July 1990, "The end of the chain: Where does decomposition of lexical knowledge lead us eventually?", (also in: *Proceedings of the 4th conference of Functional Grammar*, June 1990, Copenhagen), ESPRIT BRA-3030 ACQUILEX WP NO.010

Boguraev B., T. Briscoe and A. Copestake, August 1990, "Enjoy the Paper: Lexical Semantics via Lexicology", (also in *Proceedings of COLING*, 13th International Conference on Computational Linguistics, Vol II), ESPRIT BRA-3030 ACQUILEX WP NO.011

Copestake A., September 1990, "A system for building disambiguated taxonomies", ESPRIT BRA-3030 ACQUILEX WP NO.012

Meijs W., October 1990, "The Expanding of Lexical Universe: extracting taxonomies from machine-readable dictionaries" (in EURALEX'90 Proceedings, Bibliograf, Barcelona, 1992), ESPRIT BRA-3030 ACQUILEX WP NO.013

Calzolari N. and R. Bindi, August 1990, "Acquisition of lexical information from a large textual Italian Corpus", (also published in *Proceedings of COLING*, 13th International Conference on Computational Linguistics, Vol II), ESPRIT BRA-3030 ACQUILEX WP NO.013 bis

Working Papers of the 2nd year and a half (here enclosed)

Vossen P., January 1991, "Comparing noun-taxonomies cross-linguistically", ESPRIT BRA-3030 ACQUILEX WP NO.014

Antelmi D., Roventini A., November 1990, "Semantic Relationships within a Set of Verbal Entries in the Italian Lexical Database", in *EURALEX'90 Proceedings*, Biblograf, Barcelona, 1992, ESPRIT BRA-3030 ACQUILEX WP NO.015

Calzolari N., March 1991, "Acquiring and Representing Semantic Information in a Lexical Knowledge Base", in *Proceedings of the ACL SIGLEX Workshop on Lexical Semantics and Knowledge Representation*, J. Pustejovsky (ed.), Berkeley, California, ESPRIT BRA-3030 ACQUILEX WP NO.016

Marinai E., Peters C., Picchi E., April 1991, "A Prototype System for the Semi-automatic Sense Linking and Merging of Mono- and Bilingual LDBs", in N. Ide and S. Hockey (eds.), *Research in Humanities Computing*, OUP, ESPRIT BRA-3030 ACQUILEX WP NO.017

Marti M.A., Castellon I., October 1990, "Gramatica para el Analisis del Diccionario Vox", *Proceedings of the 6th Annual Meeting of SEPLN* (Sociedad Espanola para el Procesamiento del Lenguaje Natural), San Sebastian, Spain, ESPRIT BRA-3030 ACQUILEX WP NO.018

Castellon I.,Rigau G., Rodriguez, H., Marti M.A., Verdejo, M.F., January 1991, "Loading MRD into LDB. Characteristics of Vox Dictionary", ESPRIT BRA-3030 ACQUILEX WP NO.019

Ageno A., Cardoze S., Castellon I., Marti M.A., Rigau G., Rodriguez H., Taule M., Verdejo M.F., June 1991, "An Environment for Management and Extraction of Taxonomies from On-line Dictionaries", ESPRIT BRA-3030 ACQUILEX WP NO.020

Copestake A.A. and Briscoe E.J., June 1991, "Lexical Operations in a Unification-based Framework", in *Proceedings of ACL SIGLEX Workshop on Lexical Semantics and Knowledge Representation*, Berkeley, California, pp 88-101, ESPRIT BRA-3030 ACQUILEX WP NO.021

Briscoe E.J. and Copestake A.A., June 1991, "Sense Extensions as Lexical Rules", in *Computational approaches to non-literal language: metaphor, metonymy, idiom, speech acts, implicature*, D. Fass, E. Hinkelman and J. Martin (eds), Proceedings of the IJCAI Workshop on Computational Approaches to Non-Literal Language, Sydney, Australia, pp.12-20, ESPRIT BRA-3030 ACQUILEX WP NO.022

Montemagni S., November 1991, "Tailoring a Broad Coverage Grammar for the Analysis of Dictionary Definitions", in *Fifth Euralex International Conference*, Tampere, Finland, and in K. Jensen, G. Heidorn, S. Richardson (eds.), *Natural Language Processing: The PLNLP Approach*, Kluwer Academic Press, forthcoming, ESPRIT BRA-3030 ACQUILEX WP NO.023

Oestling A., December 1991, "Sense Extensions in the Italian Food Subset", ESPRIT BRA-3030 ACQUILEX WP NO.024

Vossen P., November 1991, "An Empirical Approach to Automatically construct a Knowledge Base from Dictionaries", in *Fifth Euralex International Conference*, Tampere, Finland, ESPRIT BRA-3030 ACQUILEX WP NO.025

Meijs W., Stemmerik Y., November 1991, "Compounds in MRDs: Analysis and Interpretation", ESPRIT BRA-3030 ACQUILEX WP NO.026

Vossen P., November 1991, "Converting data from a lexical database to a knowledge base", ESPRIT BRA-3030 ACQUILEX WP NO.027

Ageno A., Castellon I., Marti M.A., Ribas G., Rigau G., Rodriguez H., Taule M., Verdejo M.F., October 1991, "The Extraction of Semantic Information from MRDs", ESPRIT BRA-3030 ACQUILEX WP NO.028

Marinai E., Picchi E., October 1991, "A Procedure for Interactive Sense Disambiguation", ESPRIT BRA-3030 ACQUILEX WP NO.029

Amsterdam, Cambridge, Pisa groups, April 1991, "An Overview of Work on Semantic Taxonomies", ESPRIT BRA-3030 ACQUILEX WP NO.029 bis.

Ageno A., Castellon I., Marti M.A., Ribas F., Rigau G., Rodriguez H., Taule M., Verdejo M.F., November 1991, "SEISD: User Manual. Guide to the Extraction and Conversion of Taxonomies", ESPRIT BRA-3030 ACQUILEX WP NO.030

Alonge A., March 1991, "Extraction of Information on Aktionsart from Verb Definitions in machine-readable dictionaries", *in Proceedings of the Conference on Natural Language Processing and its Applications*, 11th International Workshop on Expert Systems and their Applications, Avignon, 27 - 31 May 1991, ESPRIT BRA-3030 ACQUILEX WP NO.031

Montemagni S., Vanderwende L., November 1991, "Structural Patterns versus String Patterns for Extracting Semantic Information from Dictionaries", in *COLING 14*, Nantes, forthcoming, ESPRIT BRA-3030 ACQUILEX WP NO.032

Sanfilippo A., April 1991, "Aspectual and Thematic Information in Verb Semantics", in *Belgian Journal of Linguistics*, 6, 1991, ESPRIT BRA-3030 ACQUILEX WP NO.033

Sanfilippo A., 1991, "Argument Selection and Selection Change: An Integrated Approach," in *Proceedings of the 27th Regional Meeting of the Chicago Linguistic Society*, forthcoming, ESPRIT BRA-3030 ACQUILEX WP NO.034

Sanfilippo A., Poznanksi V., April 1992, "The Acquisition of Lexical Knowledge from Combined Machine-Readable Sources," in *Proceedings of the 3rd Conference on Applied*

Natural Language Processing, Trento, Italy, ESPRIT BRA-3030 ACQUILEX WP NO.035

Copestake A., April 1992, "The ACQUILEX LKB: Representation Issues in the Semi-automatic Acquisition of Large Lexicons," in *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento, Italy, ESPRIT BRA-3030 ACQUILEX WP NO.036

Copestake A., January 1992, "The representation of group denoting nouns in a lexical knowledge base," in *Proceedings of the Second Seminar on Computational Lexical Semantics*, Toulouse, France, ESPRIT BRA-3030 ACQUILEX WP NO.037

Ageno A., Castellon I., Marti M.A., Ribas F., Rigau G., Rodriguez H., Taule M., Verdejo M.F., June 1992, "A semiautomatic process to create LKB entries", ESPRIT BRA-3030 ACQUILEX WP NO.038

Ageno A., Castellon I., Marti A., Ribas F., Rigau G., Rodriguez H., Taule M., Verdejo F., May 1992, "A production-rules approach to extracting semantic information from a lexical database", ESPRIT BRA-3030 ACQUILEXWP NO.039.

Briscoe T., A. Copestake and V. de Paiva (editors), October 1991, *Proceedings of the ACQUILEX Workshop on Default Inheritance in the Lexicon*, University of Cambridge Computer Laboratory, Technical Report No. 238, October 1991 (to be published by CUP, 1992), ESPRIT BRA-3030 ACQUILEX WP NO.040

Briscoe T., November 1991, "Lexical Issues in Natural Language Processing", in *Natural Language and Speech*, E. Klein and F. Veltman (eds.), Springer-Verlag, 1991, ESPRIT BRA-3030 ACQUILEX WP NO.041

Sanfilippo, A., T. Briscoe, A. Copestake, M.A. Marti, A. Alonge and M. Taule, January 1992, "Translation Equivalence and Lexicalization in the ACQUILEX LKB", in *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, Canada, ESPRIT BRA-3030 ACQUILEX WP NO.042

Copestake, A., B. Jones, A. Sanfilippo, H. Rodriguez, P. Vossen, S. Montemagni and E. Marinai, January 1992,"Multilingual Lexical Representation", ESPRIT BRA-3030 ACQUILEX WP NO.043

Hagman, J., December 1991, "Common and Odd Relations in Italian Dictionaries and their Treatment in Taxonomy Building", ESPRIT BRA-3030 ACQUILEX WP NO.044

Roventini, A., December 1991, "Place Taxonomies in Two Italian Machine Readable Dictionaries: General and Distinctive Features", ESPRIT BRA-3030 ACQUILEX WP NO.045

Oestling A., March 1992, "Parts and Wholes in Dictionary Definitions", ESPRIT BRA-3030 ACQUILEX WP NO.046

Hagman, J., May 1992, "Semantic Parsing of Italian Dictionary Definitions", ESPRIT BRA-3030 ACQUILEX WP NO.047

Alonge A., May 1992, "Analysing Dictionary Definitions of Motion Verbs", in *COLING 14*, Nantes, forthcoming, ESPRIT BRA-3030 ACQUILEX WP NO.048

Alonge A., May 1992, "Machine-Readable Dictionaries and Lexical Information on Verbs", in *Fifth Euralex International Conference*, Tampere, Finland, ESPRIT BRA-3030 ACQUILEX WP NO.049

Calzolari N., Hagman J., Marinai E., Montemagni S., Spanu A., May 1992, "From On-line Dictionaries to a Lexical Knowledge Base: Extracting and Representing Semantic Information", ESPRIT BRA-3030 ACQUILEX WP NO.050

Spanu A., May 1992, "Extending the Type System from the FOOD SUBSET to the PLACE SUBSET", ESPRIT BRA-3030 ACQUILEX WP NO.051

Meijs W., 1990, "Morphology and word-formation in a machine-readable dictionary: problems and possibilities", (also in *Folia Linguistica XXIV/1-2*, Mouton de Gruyter, Berlin), ESPRIT BRA-3030 ACQUILEX WP NO.052.

Meijs W., 1991, "Computers and Dictionaries", (also in C. Butler (ed.), *Computers and Written Text*, Blackwell, Oxford), ESPRIT BRA-3030 ACQUILEX WP NO.053.

Dik S., Meijs W., Vossen P., 1992, "Lexigram: a functional lexico-grammatical tool for knowledge engineering", (also in R.P. van de Riet and R.A. Meersman, (eds.), *Linguistic Instruments in Knowledge Engineering*, Elsevier Science Publishers B.V., Amsterdam, The Netherlands), ESPRIT BRA-3030 ACQUILEX WP NO.054.

Roventini A., Peters C., May 1992, "Computational Model of the Dictionary Entry: Final Report (addendum to Deliverable No.001)", ESPRIT BRA-3030 ACQUILEX WP NO.055.