TELEMATICS PROGRAMME 1991-1994

Linguistic Research & Engineering (LRE)

AN OVERVIEW

June 1994

Directorate-General XIII

Information Technologies and Industries,
and Telecommunications

Commission of the European Communities

Editors
Roberto Cencioni (DG XIII)
and
Ewan Klein (ELSNET)

Editorial assistant
Leeann Jackson-Eve (ELSNET)

Publication design
Visual Resources,
The University of Edinburgh

TELEMATICS PROGRAMME 1991-1994

Linguistic Research & Engineering (LRE)

AN OVERVIEW

June 1994

Directorate-General XIII

Information Technologies and Industries,
and Telecommunications

Commission of the European Communities

Editors

Roberto Cencioni (DG XIII)
and
Ewan Klein (ELSNET)

Editorial assistant

Leeann Jackson-Eve (ELSNET)

Publication design

Visual Resources,
The University of Edinburgh

CONTENTS

INDEX

LANGUAGE ENGINEERING

Why this technology is vital to Europe's future

Scope of this document

This document is intended to provide a broad overview of the European Commission's Linguistic Research and Engineering sub-programme within the Telematics programme (LRE in short). It places LRE within the context of Commission's related activities, in the fields of both language processing and telematics.

What is Language Processing?

Language processing encompasses a number of interrelated disciplines within the field of information technology. They share the primary goal of processing both written and spoken human language. Language processing ranges from such familiar technologies as word processing and text storage and retrieval to such ambitious pursuits as automatic translation. These technologies vary in complexity and are therefore at different stages of advancement and commercial exploitation.

In addition to the variation in complexity, there is another dimension to language processing, that of the language being handled. Because of the size of the English-speaking market and the leadership of many American companies in the information technology world, language processing is generally more advanced for the English language than for other languages, although this is slowly changing.

In short, language processing is a field with tremendous potential — both economic as well as social — but there are still substantial obstacles which prevent it from becoming more widely exploited. It may be decades before we have the listening and talking computers popularised in science fiction; moreover, it may never be possible to build a computer that can 'understand' human language perfectly. Nevertheless, the *language engineering* activities described in this document address useful tasks that can be achieved here and now.

Why Give Community Support to Language Processing?

There are compelling reasons for the Community to provide Research and Technological Development (R&TD) support for language processing within the Third Framework Programme. For one, it is part of a broad strategy to revitalise Europe's high-tech industries, which have languished in the past decade. Rather than adopt a *laissez-faire* attitude to industrial policy, the Commission is taking active steps to breath a new life into technologies which have great potential in Europe.

One of the most prominent of these technologies is language processing. This is not an accident. Historically, Europe has a rich academic tradition in linguistics. More recently, as the Member States of the European Union have grown closer together politically, with interests intertwined through trade and commerce, multilingual communication has become an intimate and indispensable part of modern life. But this poses major obstacles for companies and organisations operating beyond national borders, particularly with the advent of the Single Market.

Information and communication systems with advanced language processing capabilities have the potential to invigorate European industry by making it more productive and efficient, and thereby more competitive. In the long run, it will also help generate jobs. As perhaps the world's largest user of language technology, the Commission itself has a vested interest in encouraging this technology. Its multilingual documentation burden will only get heavier as the European Union welcomes new members from Scandinavia, Central Europe, and Eastern Europe.

However, this formidable challenge is also one of Europe's greatest opportunities, for if Europe can successfully grapple with multilinguality in all its dimensions — most importantly in terms of computer processing — then it will guarantee for itself a powerful position in the emerging global market. Success in the language engineering endeavour will be a strategic asset; the US and Japan are traditionally more monolingual cultures and may not cultivate these technologies as rapidly.

A second significant reason behind the Commission's support for language processing is that language technology can play a vital role in maintaining the plurality of European culture, where no one language should be allowed to gain ascendancy. While market forces will tend to support more widely spoken languages, the Commission is bound to counterbalance this tendency, and it must ensure that Europe's citizens have equal access to information regardless of their mother tongue. This applies to the nine official working languages of the Union as well as Europe's other languages.

A third, equally significant, reason why the Community supports language processing is because language has so much potential across the entire domain of telecommunications and information technologies, the so-called world of *telematics.* From education to consumer electronics, from office automation to special needs services, language — and therefore language processing — plays a ubiquitous role. The language processing challenge is an enabling technology which could have impact in virtually all facets of daily life.

At this point, you might ask: If language processing is so important and has so much clear commercial potential, why doesn't the Commission leave industry to exploit it in accordance with free market principles? In practice, however, even the very largest European computer companies like Siemens, Olivetti, Bull, and Philips can no longer justify the high investments involved, particularly for adapting their products to different national languages and cultures. The big American players, such as IBM and DEC, which previously made huge R&D investments in Europe, are currently wrestling with their own problems and are having to realign their priorities, as are their Japanese competitors. Because of the absence of large amounts of American-style venture capital or the Japanese-style *keiretsu* long-term investment relationships, Europe's budding language technologists are at a structural disadvantage. Hence, Community intervention is highly appropriate for this strategic field.

Community Activities

In view of both the importance of this technology and the characteristics of the European situation, the Community has been investing in language processing for nearly twenty years.

To serve its substantial internal requirements, the Commission acquired rights to the Systran machine translation system in 1975. It has continued to develop the system, adding new language pairs, and Systran is now widely used within the Commission. In the late 1970s and 1980s, the Community funded the well-known Eurotra programme, an ambitious project in which researchers from all twelve Member States built a prototype machine translation system for the nine working languages. This project created an important foundation for further work in this young field.

Within the ESPRIT programme, the Commission has supported speech and written language R&TD, both within precompetive and basic research projects. These have also helped disseminate knowledge and expertise throughout Europe.

More recently, the Community funded the LRE sub-programme, the subject of this document. The Commission is currently laying the foundation for a follow-up to these activities within the forthcoming Fourth Framework Programme (1994-1998).

Linguistic Research and Engineering

The current LRE sub-programme in the Third Framework Programme builds on results achieved in previous programmes but also reveals a shift in priorities. Whereas earlier programmes have had a very positive influence on the development of computational linguistics in Europe, particularly in southern European countries, it nonetheless became evident to the Commission that if Europe were to capitalise on language processing in the short-term, the highest priority should be given to a series of modest, tightly defined short-term goals coupled with a concerted effort to involve industry.

Inclusion of the word engineering in the name of the LRE sub-programme is a clear acknowledgement that an indispensable part of building useful applications are the time consuming, less than glamorous activities of coding data, building user and application interfaces, and undertaking cycles of iterative testing by end users. LRE was therefore born within the Third Framework Programme (1990-1994) of Community science and technology funding. It was allocated ECU 25 million for this period. In the two Calls for Proposals tendered, respectively, in 1991 and 1992, Commission officials and national experts identified three key areas for action. These were:

- general research, to tackle the many remaining research problems and foster progress from current text analysis and speech recognition techniques to more sophisticated language understanding technologies;
- common resources, tools and methods, to build over time a comprehensive infrastructure comprising databases of language data and knowledge (electronic dictionaries, computerized grammars, etc.) and the related software tools, interchange formats and pre-normative standards;
- pilot applications, to demonstrate the integration of language engineering technologies and components within information and communication systems.

General Research

The field of general linguistic research represents a natural continuation of earlier Community support for research in computational linguistics. This reflects the fact that language processing is still a relatively immature field where there are large gaps in our knowledge. Human communication is not yet well understood. While surface phenomena such as morphology (word forms) and syntax (word order) have been reasonably well codified, and progress is being made with semantics (word meaning), there are still many aspects of language which are not fully mastered; these pose hard problems in teaching machines to handle language properly.

Within LRE, several projects address specific research topics, such as discourse and semantic representation, and their results are likely to be exploited in future generations of language processing systems.

Common resources, tools and methods

When people began moving prototypes of language processing applications out of the laboratories into the field, they were dismayed to discover that it requires vast resources to build systems which can handle a wide variety of written and spoken language. These resources include extensively coded lexicons, grammatical rules, word frequency statistics, as well as raw spoken and written data for testing and evaluating systems. In particular, the compilation of lexicons is expensive (it requires skilled linguists) and time consuming (it costs many person-years). Even for English, companies find it difficult to justify the huge investments required in the area of language resources despite the huge size of the English-speaking market. For other languages, the problem is even more acute.

The Commission therefore feels that it is crucially important that it cultivates the development of shared, pre-competitive language facilities for all the official working languages of the European Union. This is very much in tune with the principle of *subsidiarity*, that is, the right action at the appropriate level of government.

These facilities include not only the above mentioned resources but also automated tools and methods for such tasks as corpora analysis and methods for evaluating language processing systems, an exceedingly difficult endeavour. This area is an obvious way in which academia and industry can fruitfully join forces; the former can supply the firm theoretical foundation, the latter can provide the data it has acquired in course of its development activities. Within LRE, a number of consortia are dedicated to just such undertakings. The end result should be resources and tools that can be reused for various applications.

Pilot applications

Developing pilot applications for various sectors is an extremely useful way of testing promising new techniques and demonstrating their feasibility and adequacy. It is also an important vehicle for involving real users in the process, for they, after all, will eventually be expected to embrace language processing technologies.

However, for this to take place, their needs must be thoroughly understood by the developers, and this implies an extensive analysis phase followed by development and testing. While this situation is no different for most industrial applications today, it is especially critical for linguistically-rich systems. Because they are by definition dealing with complex, nonfinite, and frequently ambiguous data forms — human language — such systems need to be exceptionally flexible and robust, otherwise people will simply not use them.

For these reasons, pilot language processing applications are very expensive to develop and customise, and hence the Commission considers such applications to be pre-competitive in nature. Because such projects are so costly, LRE can only support a small number of these projects. Pilot applications are a highly effective way of transferring technology from research laboratories to industry. LRE has a number of multi-disciplinary consortia developing pilot applications in such fields as medicine, law, software and aeronautics.

Accompanying measures

In addition to a series of twenty-four projects along the above lines which have resulted from the two calls for proposals, the LRE programme also includes two accompanying measures designed to provide additional support for language processing in Europe.

EAGLES

EAGLES, or more fully, the Expert Advisory Group on Language Engineering Standards, is made up of five working groups of prominent researchers and developers in the field of language processing. These groups are trying to develop some common guidelines in the areas of text corpora, computational lexicons, grammar formalisms, evaluation and assessment, and spoken language, all of which are the basic building blocks of language processing systems.

While establishing widely accepted standards is a complex matter, the EAGLES working groups can make progress defining methodologies for assessment and evaluation, and helping ensure that valuable linguistic resources are interchangeable and reusable, through the adoption of common specifications based upon widely recognised methods and practices.

ALEP

ALEP (Advanced Language Enginering Platform) is a portable development environment specifically designed for language processing. A number of research groups in academia and industry contributed to its specification and ALEP is now being developed by a consortium headed by a Belgian software house. While general purpose integrated programming environments have been commercially available for many years, ALEP is the first such system to be designed specifically with developing prototype language processing systems in mind.

This *linguistic toolbox* offers a variety of text manipulation aids, and will allow users to create and maintain lexicons, and develop grammars more efficiently. One of the primary design goals of ALEP was to make it easy to reuse linguistic data for various kinds of language processing applications. ALEP will run on standard Unix platforms, and it will be freely available throughout the European language engineering R&D community.

Summary and Conclusion

In total, the Commission has selected twenty-six projects and accompanying actions for funding within LRE, and provided some support for small scale preparatory actions in collaboration with US research bodies and Central and Eastern Europe institutes. In terms of resources, this represents ECU 22.5 million and some three thousand person-months of labour. A total of some 160 companies and organisations from throughout Europe are participating in LRE, with a significant presence of user bodies, SMEs and multinational companies.

LRE has helped forge some powerful alliances to work towards common goals and achieve tangible results. In particular, where user involvement is strong, and users themselves have a clear understanding of what they want and are prepared to cooperate to achieve it, LRE projects have a tremendous potential. The coming years should see an impressive array of tools, resources, and pilot applications emanating from LRE projects. These will serve as a invaluable foundation for further exploration, expansion and exploitation by European academia and industry, especially in synergy with related Telematics programmes.

There are some compelling reasons for the Community to support language processing. LRE and its successors in subsequent Framework Programmes will play a vital role in nurturing the field of language processing into adulthood. But language processing is more than an end in itself. Language processing should rightfully be seen in the context of telematics as a whole, for it can improve the usability and increase the functionality of applications across the entire telematics spectrum.

Additional information

For more information about the LRE sub-programme, contact:

Roberto Cencioni (programme manager)
EC, DG XIII E-4,
Bâtiment Jean Monnet
Plateau du Kirchberg
L-2920 LUXEMBOURG

Tel: +352 4301 32886

Fax: +352 4301 34999

LINGUISTIC RESEARCH & ENGINEERING (LRE)

1st Call for Proposals

Overview

1. Background

Following the Council Decisions on

(i) the third Framework Programme for Community Research and Technological Development (April 1990), and

(ii) the Specific Programme in the field of Telematic Systems in areas of General Interest (June 1991)

a first Call for proposals was launched in August 1991 for shared-cost R&TD projects for area 6 "Linguistic Research and Engineering", one of the seven areas covered by the above programme.

This Call was based on a work programme which had been adopted by the Telematics Management Committee (TMC) in July 1991 and is described in the Technical Background document.

2. Scope of the Call

The work programme for area 6 of the Specific Programme in the field of Telematics Systems identifies five lines of action. Of these key activity areas, the first Call addressed three priority areas, namely:

- Research of general interest,
- Common methods, tools and linguistic resources, and
- Pilot applications.

In particular, project proposals were invited for the following sub-areas:

Research of general interest

(a) improvement of interlinguality of linguistic representations of text/discourse;

(b) use of domain-specific knowledge to constrain linguistic interpretation of text/discourse;

(c) interfaces with speech systems and applications;

(d) use of advanced computational technologies in NLP applications.

Common methods, tools and linguistic resources

(a) advanced toolkits and working environments for language researchers, developers and professionals;

(b) methods and automated tools aiming at facilitating the reuse of existing resources and the creation of new multifunctional resources;

(c) machine-processable grammars, dictionaries, terminology collections and text corpora for the Community languages.

Pilot and demonstration projects

(a) bi- and multi-lingual machine translation;

(b) document abstracting and indexing;

(c) aids for document generation, storage, retrieval and manipulation;

(d) man-machine communication and interfaces with information and expert systems;

(e) knowledge acquisition from natural language text;

(f) computer-aided instruction, especially in the context of language learning.

3. Publication and Dissemination

The Call for proposals was published in the Official Journal of the European Communities on 21 August 1991. The deadline for receipt of proposals was 2 December 1991.

The publication of the Call was supplemented with an information package of supporting material which included, amongst others, the Technical Background Document, a document on How to Make Proposals and the Model Contract serving as a basis for shared-cost projects.

Some 800 packages were sent on request to a broad range of organisations including IT companies, SMEs, private research laboratories, universities, consultancy firms, chambers of commerce and governmental agencies. In addition, the national delegations, acting as national focal points, distributed some 300 information packages on the basis of their own distribution lists.

4. Response

The response to this first Call was very positive. By the closing date, 88 proposals had been registered. After verification of eligibility on formal grounds, seven proposals had to be rejected, leaving 81 valid proposals for technical evaluation by external reviewers.

It should be noted that the Community contribution requested by these 81 projects amounted to around 45 M ECU, thus exceeding by a factor of 6 the funds available to this Call.

5. Overview of Results

The following figures illustrate the results of the evaluation process:

- 88 proposals received,
- seven proposals rejected on grounds of formal eligibility; 81 proposals accepted for technical assessment,
- 53 proposals rejected at the end of the technical assessment phase; 28 proposals retained for in-depth evaluation of general and financial aspects,
- 13 proposals submitted to the TMC for strategic evaluation and formal opinion, of which eight were short-listed proposals and five ranked as "under consideration",
- nine proposals accepted for funding.

6. Overview of Proposals retained for Funding

The nine proposals which have been retained for funding involve some 40 participants in ten Member States, representing a wide variety of organisations, ranging from academic institutes, research laboratories and universities through SMEs to user organisations.

The individual projects, in no particular order, are as follows:

- Development of a library of linguistic data types that will facilitate linguistic description within several prominent linguistic frameworks and augment the ET6/1 formalism with a polytheoretical periphery;

- Research into a linguistic model of discourse, its representation and integration with sentence-based grammars; implementation in the ET6/1 formalism;

- Construction of computational tools for linguistically motivated corpus exploration and for dictionary population and specification management; work on descriptive linguistic specifications for reusable lexical resources and on methodological guidelines for the syntax/semantics area;

- Development of an NLP-based framework to assist in automatic indexing of technical abstracts;

- Intelligent text categorisation by means of an integrated approach involving NLP, AI and KRL methods; pilot application in the financial domain;

- Guidelines, methods and tools for software localisation and internationalisation; focus on case studies on user interfaces, help facilities and training aids for internationally marketed products;

- Creation of high-quality lexicons of European names, to be made available in machine readable form (CD-ROM) for widespread exploitation in automated systems;

- Creation of an interactive corpus-based translation drafting tool;

- Investigation into and experimentation with grammar importation, leading to a set of linguistic descriptions of "unconventional" linguistic phenomena and to methods and guidelines for grammar importation and reuse.

7. Conclusions

The first Call is to be seen as a start-up action, designed to give the necessary impetus to the development of a basic linguistic technology. The positive response to this Call shows the impact that such an action can have and demonstrates the need for a Community programme aimed at stimulating linguistic engineering through cooperative efforts of a wide range of partners from the academic world and the private sector.

Given the limited budget available to the LRE sub-programme, and in particular to this first Call, it was not expected that this Call could yield all the projects needed to fulfil the objectives of the LRE sub-programme, nor that it could result in large-scale projects involving leading business actors.

However, the fact that the Call attracted great interest from a variety of SMEs and research institutions across the EC shows that there is an increasing awareness of the economic impact of language modelling and engineering activities and a clear recognition of this field as an important element of the Community R&TD effort.

2nd Call for Proposals

Overview

1. Background

Following the Council Decision on the third Framework Programme for Community Research and Technological Development in April 19901 and on the Specific Programme in the field of Telematic Systems in areas of General Interest on 7 June 19912, a second Call for proposals was launched for shared-cost projects in Area 6 "Linguistic Research and Engineering", one of the seven areas covered by the Specific Programme.

The R&TD themes and topics relevant to the domains addressed by the call were described in technical background documents available on request from the Commission services.

2. Scope of the Call

The work programme for Area 6 of the Specific Programme in the field of Telematics Systems identifies five lines of action. Of these key activity areas, the second call for proposals issued within the framework of the LRE programme addressed three priority sub-areas, namely:

- Generic research aimed at the improvement of the scientific basis of language technologies,
- Creation of common methods, tools and language resources, and
- Pilot and demonstration applications.

A more detailed description of the topics for which project proposals were invited can be found in Annex 1.

3. Publication and Dissemination

The Call for proposals was published in the Official Journal of the European Communities on 8 October 1992. The deadline for receipt of proposals was 11 January 1993.

The publication of the proposal was supplemented with an information package of supporting material which included, amongst others, the Technical Background document, the guidelines on How to Make Proposals, a general information document outlining the rationale and background to the Community action in Area 6 and the Model Contract serving as a basis for shared-cost projects.

Around 2100 information packages were sent to a broad range of organisations (including IT companies, SMEs, private research laboratories, universities, consultancy firms and governmental agencies) based in 20 countries, either directly or through the national delegations (Telematics Management Committee, TMC) and other national focal points (e.g. science and technology agencies, regional authorities, chambers of commerce, etc.). Workshops and information days were held in Bonn, Paris, Pisa and Madrid.

4. Response

The response to this second call for proposals was very positive. By the closing date, 88 proposals had been registered. After verification of eligibility on formal grounds, six proposals had to be rejected, leaving 82 valid proposals for technical evaluation by independent reviewers.

It should be noted that these 82 proposals represent some 400 participants from 300 different organisations based in 17 countries, with a quite balanced distribution between universities (55%) and companies and R&D centres (45%). The total EC contribution requested by these 82 projects amounts to around 69 MECU (first call: 45 MECU).

5. Overview of Results

The following figures illustrate the results of the evaluation process:

- 88 proposals received,
- six proposals rejected on grounds of formal eligibility; 82 proposals accepted for technical evaluation,
- 58 project proposals rejected during the technical evaluation and, following an in-depth evaluation of management and financial aspects, 24 proposals (30%) submitted to the TMC for consideration,
- 16 proposals recommended for funding; out of these 16 project proposals, two are likely to be merged during the negotiation phase.

6. Overview of proposals retained for funding

The 16 proposals which have been recommended for funding represent around 2,000 man-months of work involving some 90 project partners[3] in all EC Member States and in several EFTA countries, representing a wide variety of organisations, ranging from academic institutes, research laboratories and universities through SMEs to IT companies and user organisations.

The individual projects, in no particular order, are as follows:

Generic research: Computational semantics

- FRACAS intends to develop a common semantic framework within which it will be possible to integrate work from different traditions of formal semantics;

Generic research: Assessment

- TSNLP aims to define a methodology for the design of diagnostic test suites for NLP systems, and to develop test suite fragments as well as an automatic test suite generation tool;
- TEMAA intends to define an assessment methodology for multilingual authoring aids, especially spelling and grammar checkers;
- SQALE intends to exploit experience gained by the proposers in the US ARPA programme, with a view to developing a multilingual evaluation paradigm for speech recognizer quality assessment;

Common methods, tools and language resources; Collaborative infrastructures

- MULTAC is to develop multilingual parallel corpora tagged according to TEI guidelines, along with the associated tools for corpus annotation, manipulation and analysis, and prototypes for extraction of multi-word terms and construction of lexica for machine translation systems;
- EUROTEXT: The objective of this project is to create an integrated set of text corpora and related tools and databases for human and machine use;
- EUROCOCOSDA intends to ensure a significant European contribution to the recently created world-wide group for speech databases and speech systems assessment (COCOSDA), and to ensure coordination at European level of several actions being undertaken within COCOSDA;
- RELATOR aims to set up and operate a prototype network of repositories of written and spoken language data, rules and tools;

Pilot and demonstration projects

(a) Machine-aided translation

- ANTHEM's aim is to develop a prototype system capable of translating medical diagnoses and to encode them into an internationally recognised classification scheme.

(b) Advanced office automation tools

- COMPASS aims to exploit machine readable versions of bilingual dictionaries to build the context sensitive look-up component of an interactive text comprehension aid for advanced language learners.
- GIST aims to combine knowledge representation and NLP techniques to develop a multilingual text generation system for instructional documents in the field of social security;
- SECC intends to develop a prototype grammar/style checker for controlled English, for use in an industrial environment.
- SIFT is to combine information retrieval and NLP techniques to build a demonstrator that will select passages in SGML texts (instructions in computer manuals) in response to a query formulated in natural language.
- CRISTAL aims to develop multilingual document retrieval facilities through a natural language interface to news articles in French, on the basis of a concept-based dictionary;
- RENOS intends to improve the performance of text retrieval techniques by building a concept network of the legal sub-language (fiscal and general legislation in several EC languages). The tools developed are to exploit both statistical and rule-based approaches.
- TRANSTERM's objective is to create a toolbox for the creation, normalisation and customisation of

terminological resources, and their integration into application specific lexicons. The tools will be validated through field tests carried out by industrial partners.

7. Conclusions

Building upon the outcome of the first call, which was a start-up action designed to give the necessary impetus to the development of a viable language technology, this second call was intended to further stimulate language engineering and to bring to maturity available results through cooperative efforts involving partners from the academic world and the private sector.

Compared to the first call for proposals, which addressed a wide range of themes, the second call has taken a more selective and integrated approach to technology development and application building, with the main emphasis on a number of focussed, complementary actions addressing a few key issues.

The call has been successful in attracting the interest of the private sector, including major IT suppliers, SMEs specialised in NLP and AI, service industries and user organisations. An interesting and unexpected feature was that several administrations and governmental bodies participated in the call.

While the size and Community dimension of the proposed projects is virtually identical to that of the first call with respect to the number of participants per project, their scope is considerably larger, with the estimated cost of 75% of the projects exceeding 1 MECU. As a result, the requested EC contribution has risen from 50 MECU in the first call to 69 MECU in this call (+38%), thus exceeding the available funds by a factor of seven.

Half of the projects liable to receive EC funds are concerned with the building of practical applications based upon existing knowledge. It is, however, worth noting that R&D tasks are present in all projects, along with integration and validation of innovative technologies.

Despite the limited budget available, the growing interest and participation of industry across the EC, including SMEs, shows that LRE has been successful in contributing to an increased awareness of the economic impact of language modelling and engineering activities in Europe and in stimulating the emergence of a European language infrastructure.

[1] Decision 90/221/EURATOM/EEC (OJ No L 117, 8.5.90, p 28-43)

[2] Decision 91/353/EEC (OJ No L 192, 16.7.91, p 18-28)

[3] This is to be compared to the projects established under the first call, which represent some 1,000 man-months of work carried out by 40 project partners.

The Advanced Language Engineering Platform (ALEP)

Objectives

Natural Language Processing (NLP) researchers and developers currently lack a solid, portable and widely accepted software platform for the development of professionally designed application prototypes. As a consequence, researchers and system developers are often forced to build the tools and development aids they need before undertaking the implementation of what matters most to them, i.e. language processing applications and related resources. This situation constitutes a major bottleneck for any serious attempt to build a viable European NLP industry.

As part of its LRE programme, the CEC has therefore undertaken the development of a generic formal and computational environment, which will be put at the disposal of EC and national R&D projects in relevant areas. By making the ALEP system widely available, the CEC intends to promote synergy between academic and industrial research centres and foster progress towards portability and re-use of research results.

Approach and Methodology

The development and distribution of ALEP is taking place in a number of stages:

- Specification and design (1991-1992)
- Prototyping (1992)
- Development (1992-mid 1994)
- Distribution (mid 1994-end 1995)

In the period under consideration (1992-1994), critical components of the ALEP linguistic tools, together with some parts of the user environment have been prototyped by P-E International. This prototype system, which is referred to as ALEP-0, embodies the core of the ALEP system and includes an easy-to-use graphical user interface and a number of basic tools for tracing, debugging and viewing linguistic objects.

The development stage comprises two development cycles. The main features of a single-user version of the system, developed by BIM during the first development cycle are:

- an architecture that is open and modular;
- a comprehensive user environment, including editors, browsers, etc., including a graphical user interface;
- linguistic processing tools based upon a unification-based formalism;
- a basic text-handling subsystem;
- an early implementation of the lexical database component..

This first version of the system, referred to as ALEP-1, was released in mid-1993 to a few pilot sites, for early assessment and feedback. The results of this assessment are now taken into account in the second development cycle.

The basic objectives of the second development cycle undertaken by BIM are:

- development of multi-user capabilities and further lingware management facilities;
- porting of the ALEP application software on to other widespread hardware platforms;
- development of a more sophisticated text-handling component;
- implementation of a more intuitive and user-friendly man-machine interface.

The result of the second development cycle, ALEP-2, will become available by the second quarter of 1994.

Exploitation and Future Prospects

From early 1994 on, the ALEP system will be used by a number of LRE sponsored RTD projects, including LS-GRAM and RGR. These projects will also contribute to the further extension and enhancement of the system. From mid-1994 on, the ALEP-2 system will be distributed to organisations that participate in EC sponsored and national projects in the NLP field. User sites will be encouraged not only to use the platform in its then current form, but to further extend and enhance it and to incorporate their preferred tools. To this effect, the CEC will provide maintenance and support services, through P-E International, until the end of 1995. These services will include training of users, the set-up of a User Group, an information hot-line, bug fixing and porting to popular prerequisite software packages.

Contact Point

Mr Paul Meylemans
Commission of the European Communities
DG XIII/E/4
Linguistic Research and Engineering
JMO Building, B4/120A
L-2920 Luxembourg

Tel:	+352 4301 32711/32886
Fax:	+352 4301 34999
e-mail:	paul_meylemans@eurokom.ie

Partners

Dr David Sedlock
BIM sa/nv
Leuvensesteenweg, 510
B-1930 Zaventem

Tel:	+32 2 719 26 11
Fax:	+32 2 725 47 83
e-mail:	david@sunbim.be

Dr Neil Simpkins
Cray Systems
11b, Boulevard Joseph II
L-1840 Luxembourg

Tel:	+352 25 08 90/91
Fax:	+352 25 08 92
e-mail:	neil@cray_systems.lu

Start Date:	January 1992
Duration:	44 months
Resources:	
Estimated cost:	2.755.000 ECU

Advanced Natural Language Interface for multilingual Text Generation in Health Care (ANTHEM)

Objectives

Most medical databases are only suited to the registration of factual knowledge without any indication about the relationship between the facts or their rationale. Only natural language provides the user with sufficient expressiveness to record diagnoses in enough detail. The ANTHEM prototype will i) transcribe the original diagnosis in a format suitable for computer processing ii) translate the diagnosis from French or Dutch into Dutch, French or German. The interface will provide the user with the ability to use a health care information system with the same flexibility as its paper-based counterpart.

The diagnoses will be encoded in an international standardised classification scheme of diseases ICD-9/10-CM. The prototype will be delivered as an Application Programming Interface for further integration in other health care systems by third party developers. A secondary aim will be to use the same sublanguage approach for the analysis of standardised medical diagnostic expressions used in disease classification systems to facilitate mapping to other systems and smooth the transition from older to newer versions (e.g., ICD-9-CM to ICD-10-CM).

Approach and Methodology

The approach is to convert the input statements into a language-independent semantic representation, which will be used as an interlingual for the translation process. SNOMED (Systematised Nomenclature of Medicine) codes, which combine seven types of medical elements (topography, morphology, aetiology, function, disease, procedure and occupation), are the basic components for this semantic representation. Beside translation, the semantic representation of a given statement will also be used to generate the relevant illness code according to the ICD-9-CM classification.

The diagnoses used as input statements are expressed in a well-limited sublanguage with a high rate of nominalisation. Thus the input is well suited for an unambiguous machine processing. Starting from medical diagnoses text corpora the medical sublanguage will be modelled and implemented in the CAT2 representation formalism. The project will use and adapt existing CAT2 lingware to build the translation modules.

The workplan includes:

1. the collation, structuring and tagging of corpora of medical diagnoses,
2. the modelling of this medical sublanguage using the interface structure of the CAT2 formalism developed during EUROTRA,
3. the representation of ICD-9-CM expressions using typed feature logic which, by means of inheritance, will support a hierarchical classification of terms,

4. the creation of a medical term lexicon in a format that makes it also accessible to other applications,

5. the development of software modules able to analyse the sublanguage input and create an abstract semantic representation, to translate it into Dutch, French and German and to generate the relevant ICD-9/10-CM code,

6. the integration of the prototype into two existing health care systems and subsequent testing in a real medical environment.

Exploitation and Future Prospects

The design of the prototype as an Application Programming Interface ensures portability and the possibility of integration in various applications. In the test phase the API will be embedded in two different systems: the one, MEDIDOC, is a PC based on-line health care information system, the other is used by the Belgian army to encode medical diagnoses in batch mode. The results will validate the approach and further adaptation of the modelled linguistic knowledge may be reused in other medical sub domains' texts.

Contact Point

Dr. W. Ceusters
RAMIT
c/o Department of Medical Informatics
University Hospital Gent
De Pintelaan 185
B 9000 Gent

Tel: + 32 92 40 34 21

Fax: + 32 92 40 34 39

e-mail: WC@FGEN.RUG.AC.BE

Partners

RAMIT VZW, Gent (coordinator) (B)

Datasoft Management NV, Oostende (B)

GFAI, Universit‰t des Saarlandes, Saarbrücken (D)

Centre de Recherche Public, Centre Universitaire, Luxembourg (L)

Ecole de Langues Vivantes, FUNDP, Namur (B)

Cellule Informatique, Hôpital Militaire de Bruxelles (B)

Universitě de Liège, Philologie Anglaise Moderne, Liège (B)

Start Date:	January 1994
Duration:	30 months
Resources:	182 person-months
Estimated cost:	1.193.540 ECU

COnstruction, augmentation and use of knowledge BAses from natural Language documenTs (COBALT)

Objectives

Managing information resources is a major issue in virtually all walks of life: the challenge is finding and manipulating useful information buried in huge amounts of text. Today this kind of text processing is usually handled in two ways: manually, or using keyword search technology. A more informed approach should use not only linguistic analysis but also elements of contextual and extra-linguistic information, that means reasoning about objects and relations in the domain of application. This basic issue is addressed by the COBALT project, which will demonstrate how different state-of-the-art language engineering technologies and pieces of software can be integrated to build a system supporting a better exploitation of information in the financial domain, through enhancements of a knowledge base. A news intelligent filtering and routing application will be built to demonstrate one of the possible uses of the tools and techniques developed within the project.

The technical goal of the project is to improve the performance of off-the-shelf text categorisation systems, by integrating current text categorisation techniques with state-of-the-art knowledge representation and selected natural language parsing and understanding techniques. COBALT will exploit existing technology, research results and software modules and concentrate R&D efforts on integration issues. The idea is thus to achieve some European innovative results on text categorisation using the results of leading technology in this field and on the state-of-the-art technology in NLP.

Euromobiliare, an independent merchant bank operating in Italy and internationally, is to be involved as application end-user. The basic language for the prototype will be English. A feasibility study for the adaptation of the prototype to other languages and other application domains is envisaged within the project.

Approach and Methodology

From a functional point of view the prototype will belong to the class of Text Categorisation systems. The basic idea is to exploit text categorisation – as performed by an existing software shell, to be reused and strengthened within the project – for a first-level category assignment and for selecting irrelevant text portions to be analysed later with natural language understanding techniques (parsing and semantic analysis). The combined results of the two analyses will enhance the original KB and thus constitute the basis for a very accurate, second-level category assignment activity.

The application domain will be represented by financial news coming from major information providers such as

Reuters, which is officially supporting the project. Research and development activity will produce an experimental "empty" system, that will constitute the basis for a real world demonstrator; the first release of this empty system will be available after twelve months and will be continuously enhanced with new software releases available during the project evolution. All the work will be done having in mind the application oriented approach of the project, and the software development will use material from the final domain from the very first stages, adopting an incremental prototype life cycle strategy. The Trading Department and the Research Group within Euromobiliare are considered likely candidates for experimenting with the Cobalt message routing system; their specific needs should determine the application requirements and constraints

Exploitation and Future Prospects

Currently there is no significant presence of European IT industry in the text categorisation technology and its application: the main products come from the USA. COBALT will start from state of the art results and extend them in terms of new technologies, greater benefits and functionality.

Experimentation in the field of online financial news intelligent routing will be carried out within the project; thus we can expect that R&D activity in COBALT will be directly exploited for the realisation of a new generation of intelligent routing applications; the basic technology, however, will be able to support the development of a large set of very interesting applications based on text categorisation. Some potential application areas are, for example, text classification, for information vending services, both automatically or interactively with domain experts, or large archives and databases intelligent navigation, for example CD-ROM navigation with simple hypertextual capabilities derived from a KB description and structuring of the CD contents.

Quinary plans to exploit the project results in two ways: a) products development; the new COBALT product on information filtering and routing will be added to the already developed Quinary product line in banking; b) technology transfer services; the project results are a natural extension of the collection of technologies Quinary is able to support in consulting, training and systems development services. UMIST will exploit results internally for its didactic activity and as a background technology for future research and development projects as in robust text processing. STEP Informatique envisages the possibility of an enhancing the Legal Advisory Systems developed in the ESPRIT II NOMOS Project (in which it is a partner) by making use of COBALT derived techniques, and is ready to take part in the creation of a commercial product of this type to be engineered from the results of the COBALT project.

Progress

After almost one year of work, considerable progress has been achieved in defining and developing the COBALT demonstrator, taking into account both technical and application constraints. The different system components have been evaluated and adapted so far to their new functionalities; they will be tightly coupled using the same underlying software. A first integrated prototype will be very soon available, to be refined in the remaining workphases basing on intensive experimentation with the user (Euromobiliare).

Start Date:	December 1992
Duration:	24 months
Resources:	132 person-months
Estimated cost:	1.010.000 ECU

Contact Point

Mr G Rocca
Quinary Spa
Via Crivelli 51/1
Milano 20121
Italy

Tel: +39 2 58 30 27 12

Fax: +39 2 58 30 53 74

Partners

Quinary SPA (coordinator) (I)

UMIST (UK)

Step Informatique (F)

Interest Groups

Euromobiliare (I)

Reuters (I)

Adapting bilingual dictionaries for on-line COMPrehension ASSistance (COMPASS)

Objectives

With texts more and more commonly on an electronic medium (e.g. word processor, CD-ROM), paper-based bilingual dictionaries are no longer as much a part of the normal working environment. The COMPASS project will evaluate, complement and convert common dictionaries to make them suitable for a computer based word and multi-word idiom translator. Moreover bilingual dictionaries have traditionally been dominated by the requirement of people either composing in the foreign language, or translating into or out of it. The emphasis of the project will be on the process of adapting existing bilingual dictionaries for foreign language comprehension, and on evaluating the user's response to comprehension tools.

The aim is to implement two bilingual dictionaries (English-French and German-English) as on-line context sensitive comprehension dictionaries. It presupposes that a user has a text on an electronic medium that he wants to read. Clicking on a word will display a context dependent translation and on request, background information (up to the full dictionary entry) in the user's mother tongue. The system will reveal if the word is part of a multi-word idiom and will select the appropriate translation depending on the syntactic context.

Approach and Methodology

The project is founded on the insight that recent advances in parsing technology may have made it possible for the look-up device itself to detect relevant features of a word's or phrase's syntactic context. At the same time, significant-sized dictionaries can now be stored in a hand-held or lap-top device. Hence this could support a display of what is being read and a context-sensitive system to look up unknown words and phrases. The system could keep a useful record of what the reader needed to look up, and hence may wish to review or memorise.

The starting points of the project are: the English-French SGML-marked machine readable Oxford-Hachette dictionary, the type-setter tape of the Oxford-Duden dictionary and a prototype called LOCOLEX that is already under development by the coordinator. The Compass project will improve this prototype through performance tuning, adding of the German-English language pair, adapting it to the specific needs of comprehending a foreign language and implementing a user interface that integrates LOCOLEX in the user's environment.

The LOCOLEX prototype carries out a morphological analysis of the sentence in which the selected word occurs and a stochastic disambiguation of the word class information. This information is then matched against the dictionary. When words with several meanings are used in a context in which there are no exploitable features that allow one to select the appropriate sense, the entry is structured as a tree and information associated with the most general node is displayed allowing the user to zoom into the appropriate sub sense.

The dictionaries will be adapted to comprehension needs by filtering out non-relevant information and many contextualising indicators, by decreasing the metalanguage and by reinforcing the treatment of the multi-word lexemes. The hierarchical structure of the dictionaries will be made explicit by transforming the source text of both dictionaries into lexical databases. The conversions starting from SGML and type-setting tape will be compared and conversion guidelines will be drawn up. Lexical gaps, missing words or collocations detected by the statistical analysis of text corpora will be filled. The human look-up process will be analysed to design a user-friendly human-computer interface.

The consortium will carry out the following actions:

1. Specification of the necessary features of bilingual comprehension dictionaries,

2. Development of methods to analyse and evaluate existing bilingual online dictionaries and to adapt them for the purpose of language comprehension,

3. Validate the methods applied to existing dictionaries, English-French (Oxford-Hachette) and for German-English (Oxford-Duden),

4. Implement a user interface that integrates LOCOLEX into the user's working environment,

5. Evaluate and test the system with users reading foreign language texts.

Exploitation and Future Prospects

The project concerns the large number of people who have some knowledge of a foreign language but not enough to read it efficiently. Since texts on electronic media are becoming more and more popular (CD-ROM, on-line newspaper, electronic mail), the number of potential users of this type of device is growing rapidly. Hence the coordinator, Xerox, may integrate a further development of the prototype in one of its commercial products.

The project will provide methods and tools aimed at facilitating the reuse of existing lexica and at creating machine-processable lexical resources. It differs from other existing projects intending to convert printed dictionaries into computer-tractable ones in the sense that the dictionaries are developed to meet a specific purpose: foreign language comprehension. Secondary results will be an improvement of the University of Töbingen German tagger and a contrastive study and encoding guidelines for two dictionaries' conversions starting from SGML and type-setter format.

The Compass consortium intends to collaborate with the EAGLES lexicon committee and will develop contacts with partners of the ACQUILEX 2 project.

Contact Point

Mrs. Annie ZAENEN
Rank Xerox Research Centre
Immeuble Le Quartz
6, chemin de Maupertuis
F 38420 Meylan

Tel: +33 76 61 50 50

Fax: +33 76 61 50 99

e-mail: annie.zaenen@xerox.fr

Partners

Rank Xerox Research Centre, Grenoble (coordinator) (F)

Institut für Arbeit und Organisation, Fraunhofer Gesellschaft, Stuttgart (D)

Seminar für Sprachwissenschaft, Universität Töbingen (D)

Department of Marketing, Advertising and Public Relations, Bournemouth Univ. (UK)

Langues Étrangéres appliquées, Université Lyon II (F)

Start Date:	March 1994
Duration:	24 months
Resources:	152 person-months
Estimated cost:	1.182.181 ECU

Survey of Language Engineering Organisations in Central and Eastern Europe and New Independent States (COOP-CEE)

Objectives

In the light of the increasing attention paid by the Community to cooperation with Central and Eastern Europe, it is essential to explore and assess the current situation in the Eastern research community in the field of language engineering, and to design scenarios for possible future cooperation between East and West, taking account of the assets and requirements manifested by the respective parties.

The outcome of this preparatory action is expected to benefit the research organisations of both the EC and the targeted Eastern countries, as it will not only provide information on potential research partners, but also voice these organisations' views on the most appropriate cooperation themes and schemes. At the same time, it is hoped that this action provides background material and decision aids for use by potential sponsors, including the CEC, in the preparation and implementation of forthcoming international cooperation actions.

The project is launched and funded from within the EC COPERNICUS line of action.

Approach and Methodology

The European Network in Language and Speech (ELSNET) is undertaking a survey of organisations in Language Engineering (including Natural Language Processing and Speech Technology) in Central and Eastern Europe and selected New Independent States (C&EE/NIS). The goals of the survey are twofold:

Fact-finding

To establish a reliable, up-to-date picture of organisations and their activities in C&EE/NIS, across three dimensions. The first dimension is geographical; the second is organisational, covering the span from academic and industrial research laboratories through professional associations to government agencies; the third is sectoral, comprising R&D in natural language processing and speech technology, standardisation, dissemination, training, and policy making.

Synthesis

To evaluate the strengths and weaknesses of C&EE/NIS expertise and infrastructure; to assess the market potential of their language technology; to assess the potential for greater coordination within the C&EE/NIS region and for collaboration between the research communities in C&EE/NIS and the European Union; to recommend models for cooperation at a variety of levels.

Benefits and Future Prospects

One important effect of the survey will be to stimulate awareness of the problems and potential of language engineering in the C&EE/NIS region. At the moment, information about the current state of affairs is hard to find, dispersed and partial. The fact-finding component of the survey will integrate existing information sources, gather new facts as required, and the results will be widely disseminated in both paper and electronic formats. In addition, the data to be published in the survey will be a valuable source for R&D teams wishing to identify C&EE/NIS partners for projects in research and industrial development.

The synthesis component of the survey will help initiate a wide-ranging discussion of priorities, objectives and resource implications for future cooperation actions at both the national and European Union level.

It will present a number of options for collaboration, and is expected to become an important source of facts and models for Western policy-making and funding agencies. The synthesis will be arrived at on the basis of discussion with a wide and representative range of informed actors, including the ELSNET sites, professional associations like EACL and ESCA, and prominent figures from the Eastern R&D communities.

Coordinator

Dr Ewan Klein
University of Edinburgh

Contact Point

Ms Dawn Griesbach
Centre for Cognitive Science
2 Buccleuch Place
Edinburgh EH8 9LN

Tel: +44 31 650 4594

Fax: +44 31 650 4587

e-mail: elsnet@cogsci.edinburgh.ac.uk

Associated Partners

CNRS/LIMSI (FR)

Univ. Saarland (DE)

CNR/ICL, Univ. Pisa (IT)

Start Date:	January 1994
Duration:	six months
Estimated cost:	64.670 ECU

International Cooperation for EAGLES (COOP-USA)

Background and Objectives

The overall aim of this LRE action, which is funded through DG XIII's special budget for international Science and Technology cooperation, is to stimulate and support cooperation between European projects and initiatives and corresponding North-American activities, in the field of Language Engineering and Resources. In this area there has been a sweeping upsurge of activity in the last few years, in both continents, with regard to the gathering and processing of sizeable language resources as well as to their standardisation, to render them usable for a wide range of systems and applications.

Following extensive discussions between representatives of US sponsoring agencies and the CEC, also involving coordinators of primary US and European research networks, projects and initiatives, a number of small-scale actions have been initiated (see below), primarily addressing state-of-the-art and pre-normative standardisation, and creation and exchange of multilingual language datasets.

From the European point of view, the aim is to continue and/or extend relevant on-going and about-to-start LRE projects, such as EAGLES (61-100), RELATOR (62-056) and EURO-COCOSDA (62-057), by providing support for a co-ordinated European participation in major international activities. The activities funded under the present project are intended as a start-up action to lay the groundwork for broader transatlantic cooperation in the future, which may at a later stage be extended to encompass other critical cross-sectoral activities such as the shared development of large-scale pre-competitive language resources and the joint definition of methods and techniques for the evaluation and benchmarking of language processing systems.

Actions supported

Survey of the state-of-the-art in natural language and speech processing

The objective of this action is to survey the current level of accomplishments in specific NL and Speech areas relevant for the empirical verification and practical impact of the effectiveness of language and speech processing. A further objective is the identification of the most promising developments over the next three to five years in NL understanding, processing and modelling, and Speech analysis, synthesis and recognition.

This study is to provide a reasonably detailed analysis of

i) available R&D results and technologies, and

ii) R&D directions, for use by research institutions, industrial laboratories and funding agencies.

The sponsors of this initiative (DG XIII/E and US NSF) have assumed the overall responsibility for the preparation of the study report as general editors. Six world-level specialists, three from the EC and three from the US, have been invited to serve on the editorial board. They will carry the main scientific and organisational responsibility for the study.

The report is scheduled for publication in late 1994.

Standardisation of textual and lexical data

This action is designed to support the participation of European institutions and individuals in the Text Encoding Initiative (TEI), with regard to the 1993-94 operations.

The TEI has been initiated and jointly sponsored by the Association for Computational Linguistics (ACL), the Association for Literary and Linguistic Computing (ALLC) and the Association for Computers and Humanities (ACH). TEI has received substantial grants from the US National Endowment for the Humanities, the Andrew W. Mellon Foundation and the Social Science and Humanities Research Council of Canada. Partial support has also come from DG-XIII. The goal of the TEI is to provide publicly defined guidelines for a common interchange format, in which computational linguists, humanities researchers, librarians, application developers and information technologists can encode and exchange textual and lexical data in machine-readable form.

More specifically, in 1993-1994 the TEI intends to achieve the following objectives:

1. Complete the second version of the Guidelines, including the preparation of tutorials and casebook materials, and arrange for their publication and electronic distribution;
2. Conduct workshops and tutorials, in the USA as well as in Europe; prepare a set of manuals for specific categories of users;
3. Establish work groups for new areas, including: TEI tags in the context of information retrieval, alignment mechanisms for multilingual corpora, alignment mechanisms for coordinating speech with speech transcriptions, and possibly others;
4. Carry out evaluation and usability studies and investigate related software development;
5. Establish more formal relations with other relevant organisations, including ISO, ARL and Infoterm; and
6. Prepare a detailed proposal for the transformation of the TEI into a permanent international structure.

Multilingual corpus collection

The MLCC project aims to deliver two corpora, one comparable polylingual, loosely modelled on the TREC/TIPSTER corpus, and one parallel multilingual, derived from official CEC documents. The two corpora should serve international cooperation providing:

1. a document collection with connections to the TREC/TIPSTER materials in the US, enabling compatibility in research

in Information Retrieval (IR) and NL system evaluation; and

2. a basis for exchange of textual data with parallel American initiatives (especially LDC and ACL/DCI), and for cooperation in developing methods for corpus annotation (e.g. tree banks).

Furthermore, the texts collected will also provide a test-bed for several EAGLES activities: in particular to test the specifications for text representation, text annotation, multilingual text alignment, and to develop and possibly test criteria and methodologies for corpus based evaluation and assessment of NL and Speech systems.

The goal of this effort, which is coordinated by the University of Edinburgh, is to publish the polylingual and multilingual corpora on CD-ROM in late 1994, and to distribute them at cost through ELSNET and the LRE project RELATOR in Europe, and through the LDC in the US.

Contact Point

Prof A. Zampolli
Dipartimento di Linguistica
Computazionale
University of Pisa
Via della Faggiola 32
I-56100 Pisa

Tel: +39 50 56 04 81

Fax: +39 50 58 90 55

e-mail: eagles@icnucevm.cnuce.cnr.it

Start Date:	January 1993
Duration:	20 months
Estimated cost:	195.000 ECU

Conceptual Retrieval of Information using Semantic dicTionAry in three Languages (CRISTAL)

Objectives

The project addresses the area of text indexing and retrieval, its goal being to provide access to textual information by matching query and text concepts rather than by string or keyword matching. Thus the user will be given the ability to search for an idea, without requiring knowledge of the texts examined or mastery of a cryptic query language. The CRISTAL project will develop a multilingual (French, English and Italian) natural language interface in order to retrieve monolingual (French) text in a corpus of newspaper articles.

The system will integrate linguistic methods and information retrieval techniques. It will reuse many existing components: a conceptual dictionary Dicologique, as well as results from other CEC sponsored projects: SIMPR, PLUS and COBALT.

Approach and Methodology

The project will be carried out along two development axes.

1. The adaptation of the French conceptual dictionary, Dicologique, a device that maps natural language lexemes into concepts. This involves an expansion of the structure to accommodate multilinguism (English and Italian) and the semantic analysis of English and Italian subsets. The necessary software tools to consult and update the conceptual dictionary will be built during the project.
2. The development of a concept-based information retrieval environment that includes an indexing module, a search engine and a dialogue management module. The interface will first accept a natural language query and then refine and disambiguate this query through a dialogue with the user. Concept based retrieval will be investigated.

The consortium is composed of industrial partners, research organisations and a user. The user will provide real corpora and will participate in the requirement specification. The project will demonstrate and evaluate the techniques and tools through final end-user applications.

The conceptual dictionary Dicologique is composed of a multi-hierarchical tree structure where each leaf corresponds to a word or a phrase. A variety of types are used to characterise the nodes. These types enable grouping of concepts by topics, building of links (IS_A, SORT_OF, PART_OF...), linking near-synonyms, and grouping concepts by characteristics (size, shape, etc).

The conceptual dictionary will be extended to multilinguism by giving English and Italian equivalents for each concept of the studied subset, on a 1 to 1 basis. Where exact word equivalents are not available, phrases will be used. These direct links between words will avoid creating three isomorphic semantic structures for the three languages used in the query input.

The parser will comprise a morphological and a syntactic analyser and an interpretation component. The main purpose of the parsing is to disambiguate syntactic senses of words in the document texts as well as in the natural language queries. The interpretation component of the parser will then map the syntactical output to the conceptual dictionary. The approach to solving ambiguities will rely whenever possible on the context of the document or the context of the query conversation, otherwise it will be solved by questioning the user. The dialogue manager will be simplified compared to other systems through constraining the expected user response. The Esprit project PLUS demonstrator is the starting point of the dialogue module.

To enable multilingual access to the text database, the documents will be indexed monolingually but queries will be processed multilingually. The concepts extracted from the query are substituted by their target (French) equivalents, which are then used in the indexing process. A formal notion of semantic distance will be defined during the project and a threshold will enable too distant concepts in the matching process to be filtered out.

Exploitation and Future Prospects

The project expects scientific results in the fields of man-machine communication, dialogue management and conceptual dictionary building. It will test the theoretical models developed during the previous years in a concrete commercial domain. Cooperation is foreseen with other LRE indexing and information retrieval projects.

The major improvement compared to off-the-shelf products results from the combination of:

1. multilingualism; the user is able to access information in a foreign language without needing a perfect knowledge of that language,
2. the ability to access information in free natural language, and
3. the ability to search for an idea as opposed to keyword matching.

A follow-up of the project could turn the prototype into a commercial product. The project aims at a generic application that provides electronic access to current information. Access to remote data-banks over network services such as Minitel "information kiosks" and direct access to bulk data distributed on CD-ROM are potential applications. The approach is domain independent and the system could also be adapted for public information suppliers or for engineering purposes (technical documentation, maintenance manuals, test reports, etc).

Contact Point

Mrs. Krystyna Laus-Maczynska
Cap Gemini Innovation
86-90, rue Thiers
F 92513 Boulogne-Billancourt Cedex

Tel:	+33 1 49 10 52 82
Fax:	+33 1 49 10 06 15
e-mail:	laus@capsogeti.fr

Partners

Cap Gemini Innovation, Boulogne-Billancourt (coordinator) (F)

Cap Volmac, Rijswijk (NL)

Institute for Science and Technology, University of Manchester (UK)

Istituto di Linguistica Computazionale, CNR, Pisa (I)

Memodata, Caen (F)

Associated Partner

L'Europèenne de Donnèes, Boulogne (F)

Start Date:	December 1993
Duration:	30 months
Resources:	149 person-months
Estimated cost:	1.612.250 ECU

61-034

Descriptive Lexical Specifications and Tools for Corpus-based Lexicon Building (DELIS)

Objectives

DELIS is a multidisciplinary project with three broad objectives:- to contribute to a methodology of dictionary development based on corpus evidence; to produce parallel dictionary fragments in five languages, and to produce software tools supporting this kind of lexicographic work.

Its methodological goal is to use syntactic phenomena found in corpus evidence to define properties of lexical semantic classes, individual lexemes belonging to these classes and the readings of such items. Its descriptive goal is to produce a set of parallel dictionary fragments for English, French, Italian, Danish and Dutch, covering selected lexical semantic classes. In parallel with this work, software tools will be specified, implemented and integrated in a common user environment, providing computational support for the lexicographic work and the underlying methodology. These tools will include tools for corpus exploration and for the manual acquisition of lexical knowledge, its management and the "population" of previously defined typed feature based models and its eventual (SGML-based) exportation and presentation in dictionary articles.

Approach and Methodology

DELIS is a concrete, albeit incomplete, example of corpus-based design of multi-functional dictionaries as developed and discussed in the Eurotra-7 study. It is based on the assumptions that:

- the criteria according to which lexical items are classified must be made as explicit, communicable and thus reproducible as possible by binding them to pieces of observable linguistic phenomena;

- a single representation formalism, adequately supported by computational tools, leading to a consistent descriptive specification is required in order to use corpus evidence as a raw material for the linguistic description of lexical items (TFL, as an emerging standard, will be used for this purpose);

- tools designed for the handling of descriptive linguistic specifications need to be generic with respect to the linguistic "container" (i.e. independent of the "contents"), but they must also accommodate initial user requirements and subsequently be tailored according to the results of live testing.

The project software will be produced with the assistance of professional users from a dictionary publishing house and a translation/documentation company, in the form of requirements definition, feedback on specifications and field tests of early prototypes.

Exploitation and Future Prospects

DELIS is an interdisciplinary technology-transfer project, making technologies which have been developed and are now beginning to be used in NLP available for lexicographic work in translation/documentation and publishing. It will also make a significant contribution to the research areas of linguistic (particularly semantic) description and the integration of typed feature systems and user interfaces.

In particular, DELIS will contribute to a methodology of structuring semantic and syntactic information so that it is independent of editorial tools used to manage formal, typographical and other characteristics of lexical information. Ultimately the creation of product-independent lexical databases that can be used for more than just traditional paper dictionaries is envisaged.

The DELIS prototype will be parameterizable and thus adaptable to the systems and databases used by the various project participants.

Progress

Achievements so far include surveys of linguistic and computational approaches relevant to DELIS work, an outline of the architecture of DELIS lexicons and of the DELIS toolbox as well as a definition of methods and scenarios for field validation and a specification of users' requirements. Draft versions of a formal linguistic model, the "corpus evidence encoding schema for DELIS" (CEES), and of the toolbox specifications are also available and should be finalised and delivered by the end of November 1993.

Contact Point

Ulrich Heid
Institut für maschinelle
Sprachverarbeitung,
Computerlinguistik
Universität Stuttgart
Azenbergstrasse 12
D-W-70174 Stuttgart 1

Tel: +49 711 121 1373

Fax: +49 711 121 1366

Partners

Univ. of Stuttgart, Inst. Masch. Sprachverarbeitung (coordinator) (D)

Sonovision Itep Technologies (F)

Van Dale Lexicografie b.v. (NL)

Istituto di Linguistica Computazionale del CNR (I)

Center for Sprogteknologi (CST), Copenhagen (DK)

Associated Partners

Lingsoft Inc. (FI)

Linguacubun Ltd. (UK)

Vrije Univ. Amsterdam (NL)

Start Date:	February 1993
Duration:	26 months
Resources:	170 person-months
Estimated cost:	1.350.000 ECU

61-062

Towards a declarative theory of Discourse (DISCOURSE)

Objectives

Many of the problems the current, sentence-based NLP systems are confronted with are due to the lack of an adequate theory of discourse. The interpretation of linguistic expressions is often determined by extra-sentential information, and systems that do not take such information into account will almost inevitably perform inadequately. Most treatments of discourse phenomena (e.g. pronoun resolvers) follow a procedural methodology which makes them uninvertible and hence unsuitable for integration into an NLP system. This project aims at a declarative theory of discourse, a grammatical approach to pronouns, temporal anaphora and discourse relations.

Approach and Methodology

The project will be based on selected existing theories of discourse, and will build on these to arrive at a declarative theory of discourse for NLP, to be implemented in the ALEP environment. The ALEP platform is an environment for lingware development and testing being developed by BIM with Community funding. Another EC-funded project whose results will greatly benefit the present project is ET10/61 'Formal Semantics for Discourse'.

The project is divided into five clusters of work packages, each with its own objective. A discourse analysis package will produce linguistic specifications of discourse relations concentrating on nominal anaphora, temporal anaphora and sentence relations. A work cluster on representation will specify the representation of discourse information in a sign-based grammar. A computation package will investigate the ways in which discourse information can be computed using principles such as declarativity, reversibility and compositionality, and how these can be reconciled with the theory developed in previous work packages. The implementation work package will result in a partial implementation of the discourse model in the ALEP system. Evaluation, reusability, comparisons with other work and the relevance of the project to the NLP community will all receive careful attention.

Exploitation and Future Prospects

The project is expected to make significant contributions to the following state-of-the-art interfaces:

- anaphora resolution and declarativity
- discourse and sign-based representation
- sign-based representation and generation.

Results will be made accessible to a wide range of grammar writers and computational and theoretical linguists and be presented at conferences and workshops.

Progress

The work undertaken during the first project phase has resulted in a preliminary specification of a linguistic model for discourse grammar, temporal anaphora and nominal anaphora. Work is proceeding on schedule on the formalisation of the selected discourse phenomena, the definition of requirements for system and grammar, the design and implementation of a fragment of sign-based sentence rules ("backbone grammar") and the specification of a sign-based representation for (part of) the discourse fragment.

An extended, refined and integrated version of the linguistic model for discourse and a preliminary account of the specification of the discourse grammar in a sign-based formalism will be delivered by the end of 1993.

Contact Point

Joke Dorrepaal
Stichting Taaltechnologie
Trans 10
NL-3512 JK Utrecht
The Netherlands

Tel: +31 30 536 062

Fax: +31 30 536 000

Partners

Stichting Taaltechnologie (coordinator) (NL)

University of Edinburgh, HCRC (UK)

University of Amsterdam (NL)

Université de Clermont-Ferrand (F)

Start Date:	January 1993
Duration:	25 months
Resources:	103,5 person-months
Estimated cost:	710.000 ECU

Expert Advisory Group on Language Engineering Standards (EAGLES)

Objectives

The integration of natural language and speech processing into complex Information Technology applications has been hampered by the lack of generic technologies and of large-scale language resources. In Europe, there is particular concern as the Language Industries are mainly driven by SMEs providing highly customised applications. These are slow to develop and maintain, mainly because of the high costs of building the natural language and speech resources required for such applications. An associated problem is the diversity of formats and variable linguistic specificity of existing resources which hinder their reuse and engender duplication of effort. Several European projects and interest groups have been addressing these problems, and it has become clear that a more systematic approach is needed to achieve consensus for both linguistic information and its representation.

The EAGLES initiative aims to establish a set of coordinated expert groups ('the Group') in the area of pre-normative linguistic research. Counting on the active collaboration of more than 30 research centres, industrial organisations, professional associations and research networks across the EC, the Group is designed to become a driving force in the process of harmonisation of methods for the creation, description, representation, evaluation and assessment of linguistic data in both the natural language and speech field. However, as Language Engineering is still in its infancy with regard to prominent Information Technology sectors, it is difficult and possibly counter-productive to attempt to specify and enforce standards in the short term. It is therefore the aim of this project to gather and assess existing methods and to make recommendations as to what is judged to be the best currently available practices.

In its later stages, the project will also contribute to the demonstration, validation, promotion and dissemination of its achievements, including making them publicly available, in close cooperation with its associated European R&TD projects and relevant infrastructural measures and professional associations (e.g. RELATOR, ELSNET and ESCA). In a longer-term perspective, it is hoped that the growing acceptance of EAGLES results by the R&TD communities will enable the Group to submit standards proposals to the national and European standardisation bodies active in relevant fields.

Approach and Methodology

The Group is operated through three organisational instances:

- Management Board,
- Working Groups and
- Hosting Organisations,

jointly supported by the Coordinator's secretariat and editorial board.

The Management Board comprises the representatives of several European project consortia established under the EUREKA, ESPRIT and LRE programmes, that are primarily engaged in the design and building of linguistic resources or are 'consuming' such resources (e.g. application developers). Other participants are representatives of the European associations and coordinating bodies ESCA, ELSNET, FOLLI and the European Chapter of ACL. It is the task of the Management Board to ensure the dissemination and endorsement throughout these participating parties, and in affiliated projects, of the results produced by the Working Groups. It will also ensure the proper functioning of the other bodies in the Group and adjust, where necessary, the orientations, work plans and procedures of the Working Groups to emerging requirements. As the project goes on, additional bodies, preferably from industry and corporate user groups, are expected to join the Board.

Working Groups, each supported by a Hosting Organisation and represented by a chairperson on the Management Board, have a specific field of activity. Five of these have been established to date: Text Corpora, Computational Lexica, Formalisms, Evaluation and Spoken Language resources and methods. A Central Editorial Board is responsible for the elaboration of reports and other deliverables. Each group creates suitable working structures and procedures in cooperation with the Hosting Organisation in order to carry out the tasks stipulated in the work programme developed by the Management Board.

Exploitation and Future Prospects

Producing standardised, reusable resources can bring benefits both in terms of cost savings, since a wide range of 'lingware consumers' can access and use them, and in terms of a reduction in the variety of software tools required to manipulate them. Companies can increase their competitiveness, as reusable resources will provide the basis from which application-specific resources can be derived, while the academic world will profit from the ability to share resources for further R&TD activities.

The development of pan-European standards in language engineering is a long-term endeavour. By stimulating the coordination of effort, consensus building and early exchange of results, the EAGLES initiative aims to accelerate the process and bring this goal closer.

Progress

Structure and composition of the Group's functional bodies have been consolidated and all work plans for the individual Working Groups have been defined and adopted. The consortium aims to complete its first stage of operations by mid-1994 with the provision of a preliminary report for which feedback will be invited from the associated parties and other relevant bodies. The overall results stemming from the evaluation of the report and additional work carried out in a second work phase are expected to be made widely available by mid-1995, in the form of a handbook.

Contact Point

Prof. Antonio Zampolli (Sec. Tarina Ayazi)
ICL/CNR
32, via della Faggiola
I-56100 PISA

Tel: +39 50 56 04 81

Fax: +39 50 58 90 55

e-mail: eagles@icnucevm.cnuce.cnr.it

Partners

Consorzio Pisa Ricerche, Pisa (coordinator) (I)

GSI-ERLI, Paris (F)

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Saarbrücken (D)

Center for Sprogteknologi (CST), Copenhagen (DK)

Vocalis Ltd., Cambridge (UK)

Instituto Cervantes, Madrid (E)

Start Date:	February 1993
Duration:	30 months
Estimated cost:	1.250.000 ECU

62-057

EUROpean interface to COCOSDA (EUROCOCOSDA)

Objectives

The development of common and widely accepted evaluation methods in the areas of language and speech technology – and of suitable resources for implementing these methods, such as spoken and written language databases – is an international cooperative effort where Europe should play a crucial role. In order to give a concerted contribution to this global effort, the project aims to support and coordinate European participation in COCOSDA – the International Coordinating Committee on Speech Databases and Speech Input/Output Systems Assessement. This recent world-level development in language and speech engineering – with representatives from about twenty contries, drawn from Europe, North America, China, Japan and the Pacific Rim area – is concerned with the definition and application of multi-language databases and assessement standards and protocols in the field of Spoken Language Engineering. It is currently based to an important degree on prior European work; there is, however, no organised European presence in its activities.

The following objectives are consequently at the core of the EuroCocosda project: give a focus for European work within, and input to, the new world frame provided by Cocosda; contribute to the internationally recognized need for a synergy between the areas of natural language and speech; provide for the joint production of spoken and written language databases, in order to support cooperation among these areas.

Approach and Methodology

Different action lines will be implemented within the project in order to reach these aims. A unifying infrastructure will be established – EuroCocosda – for concerted European contribution to the main Cocosda world group. Within this framework, direct links will be maintained both with the Cocosda Central Commettee and the three Cocosda Working Groups: Recognition, Synthesis and Corpora. An organisational support foundation for the European component of a world wide telephone speech database – Polyphone – will be provided; the funding for each individual language's 5000 speakers corpus should come from national PTT resources. A coherent framework will be created both for inputs to the Cocosda world speech synthesis database and for the European component of the NL/Speech corpora initiative, NEWS, a multilanguage newspaper text and speech database.

The TED corpus (Translanguage English Database), containing spontaneous speech, read speech and some associated text material, will be directly acquired and and formatted within the project. Two groups of recorded speakers – native speakers of English with different dialects and non-native speakers of English as a foreign language – will provide multi-dialectal and multi-accent speech data. Speakers will be recorded at international conferences on speech communication; the associated text material will be organised and structured in a distributable form and made available to the scientific community at an early stage, thus providing an excellent link between natural language and speech technology. The issue of reusability will be also specifically focused in this data collection effort.

Links between COCOSDA and relevant European and international initiatives (LRE projects SQALE, RELATOR, EAGLES; ELSNET; LDC) will be created or fostered, thus providing an official and regular communication channel for exchanging experiences and data. Present needs of the scientific and user community will be surveyed and future initiatives prepared.

Exploitation and Future Prospects

The following benefits and results are foreseen for the project, with special reference to the access from Europe to developments on the world scene and with a substantially enhanced possibility for European norms and de facto standards to become more widely influential and accepted:

- A central position for Europe in the international scene, with major initiatives now capable of coming from EC countries acting in concert.
- Better coordination between European laboratories, and a world level dimension for the joint activity of European NL and speech based workers.
- An insight into world – in particular US and Japanese – technical expertise, coming from direct collaboration on common projects .
- The possibility of harmonising national with world actions on database recording and labelling.
- The availability of poly-language databases for developing and testing telephone based speech applications in both recognition and synthesis.
- The availability of world relevant database frameworks for developing and testing multilingual speech synthesis systems.
- A survey on the existence and availability of newspaper databases in various languages; the potential to create and influence the creation of new ones at the international level.
- A domain specific multi-language and multi-accent corpus framework, with a large number of speakers, speaking the same language (English) in a natural way, under moderate stress, during a relative large amount of time. The corresponding text material will allow the building of lexica and language models.

Although the project is concrete and small-scale, primarily focused on the precise objective of an urgent need for European coordination within the framework of an international body which already exists, it will however prepare the ground for larger scale operations, whilst stimulating cooperation at the European and international levels.

Contact Point

Prof. Adrian Fourcin
Dept. of Phonetics and Linguistics
University College London
Wolfson House, 4 Stephenson Way
London NW1 2HE
United Kingdom

Tel: +44 71 380 7401

Fax: +44 71 383 0752

e-mail: euro@phon.ucl.ac.uk

Partners

University College London (Coordinator) (UK)

LIMSI -CNRS (F)

University of Amsterdam (IFA) (NL)

CSELT (I)

University of Munich (D)

Start Date:	December 1993
Duration:	20 months
Resources:	29 man/months
Estimated cost:	190.700 ECU

A Framework for Computational Semantics (FraCaS)

Objectives

Future man-machine communication will be multi-modal, including visual, tactile and acoustic interaction. Dialogue between humans and machines will be a crucial component of multi-modal communication. Currently, free dialogue between humans and machines is not feasible, mainly because computing the meaning of unrestricted utterances is a long-term goal which requires further research.

FraCaS will address this problem along two interrelated axes: (a) present a unified view of currently available results with special emphasis on their descriptive and computational adequacy, and, more importantly, (b) put future research efforts in this area on better grounds, avoiding to a large extent the currently unavoidable duplications of efforts.

The aim of the project is to bring about a substantial convergence in computational semantics with the benefit of a substantial saving of duplicate work. In particular, the project will:

1. present an informal framework which allows comparison of current semantic approaches both with respect to their claims and their usefulness for implementation;
2. present the main semantic approaches in terms of this framework;
3. examine the feasibility of a general computational framework;
4. make preliminary investigations of the formal specifications for such a framework;
5. apply the framework to a representative fragment of real-life language;
6. draw together the results of consultation with a representative base of researchers in the field.

Approach and Methodology

The work comprises a survey of the main approaches to computational semantics, a comparison of these approaches with respect to a number of criteria central to the field, a characterisation of current approaches in term of a common framework and the evaluation of their computational adequacy. The work will be carried out in co-operation with an extended panel of experts representing all major centres active in the field in order to ensure representativeness of the work to be carried out.

Exploitation and Future Prospects

The results of FraCaS will contribute substantially in putting future developments, be they of practical or theoretical nature, onto firmer ground while guaranteeing maximal computational usefulness. They will also encourage the development of guidelines and standards for representation and interchange of semantic information.

Contact Point

Dr Robin Cooper
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
Scotland

Tel: +44 31 650 44 07

Fax: +44 31 650 45 87

e-mail: cooper@cogsci.ed.ac.uk

Partners

University of Edinburgh (UK)

CWI, Amsterdam (NL)

Universität des Saarlandes (D)

Universität Stuttgart (D)

Start Date:	January 1994
Duration:	25 months
Resources:	130 person-months
Estimated cost:	975.000 ECU

62-009

Generating InStructional Text (GIST)

Objectives

The efficient production and handling of multilingual documents of various nature is one of the major challenges to be faced in the short term by international organisations, public services and industry in Europe. The GIST project intends to address precisely this challenge by developing a multilingual generation system for texts describing bureaucratic procedures (e.g. how to apply for a social security facility, what to do in order to get visas, etc.), starting from language independent specifications and making use of advanced NLP nad knowledge representation techniques. Three languages will be taken into consideration: English, German and Italian. The final prototype is expected to provide good quality drafts of texts; such drafts are then to be revised and post-edited by professional writers and/or translators.

The system is meant to significantly shorten and improve the process of production of instructional texts. Also, the possibility of storing and reusing previously edited, language independent descriptions will improve the effectiveness of the GIST system. The prototype will constitute the starting point for a generation of systems for supporting the work on procedures in Public Administrations and in large companies. The idea is to embody the drafting functionality in an integrated system with several other functionalities for storing, retrieving and analysing structured representations of procedures.

Approach and Methodology

A language independent specification will be provided by a user by means of a graphical interface, built on top of the knowledge representation system where the relevant domain knowledge is stored. The interface will feature facilities for assessing the consistency of the description, and for its storage and retrieval. The (language independent) description serves as input for a module that – for each language – produces a text plan which respects language- and domain-specific requirements. The text plan is translated into the format of a Sentence Plan Language, which serves as input specification for the tactical generators

The project is largely based on the re-use of existing technologies, mainly developed by the partners in the framework of national and international research programmes. Other than this, the development will be based on extensive empirical research. Empirical investigations will address the occurrence and interdependence of discourse phenomena in the selected corpus of texts. They will also deal with the determination of needs of the text producers (domain experts, technical writers and translators) in their everyday work. These examinations are made with respect to the three languages: English, German and Italian. The results of the empirical investigations will be stated in formal terms and will be represented in knowledge sources. Another issue dealt with in the project is the identification of textual phenomena which are common to various languages, allowing a common representation. Resource sharing is emphasized and redundancy avoided.

For the development of the final drafter prototype, the project is divided into the phases of requirements analysis, adaptation of tools, theoretical specifications and practical implementation of the drafter components, integration and evaluation. Users will be participating in various stages of the project by defining assessment criteria and evaluating prototypes developed throughout the project.

Exploitation and Future Prospects

The development of the GIST system advances the state-of-the-art with respect to the following points:

- *multilinguality:* different text-linguistic means can be used to express the same message in various language; these findings are encoded in the design of the GIST drafter, which produces texts which account for language-specific differences..
- *interrelationships between various textual phenomena:* the project examines and formalizes the interactions between text structure and the textual phenomena of anaphoric reference and thematic progression. They are represented in the system by means of language-dependent and language-independent components.
- *user participation:* the potential users are involved starting from the first phases of the prototype development. The needs of technical writers and translators are therefore encoded in the system.
- *reusable knowledge sources:* the knowledge sources developed in the course of the project will be represented in a modular and declarative way; this method facilitates the extension and adaptation of the prototype to new applications and new domains and also makes the components reusable for other NLP systems, like text understanding, machine translation or automatic abstracting systems.
- *relevant industrial application of prototype:* the drafter prototype can be directly applied and easily extended to the generation of multilingual procedural texts as produced in Public Administration and large companies daily.

The use of the GIST system by the selected user groups (INPS, PAB) will have an important role in showing the benefits of applied Natural Language Processing to other users, especially Public Administrations in multilingual areas. The industrial partner (Quinary) will use the prototype to develop a marketable product for the drafting of multilingual texts, and promote it in areas concerned with multilingual documentation; new application areas and- functionalities will be explored, as well as the integration of the prototype – or its subsequent developments – into more complex systems. The partners with a background in applied research (IRST, ITRI, OFAI) will be responsible for promoting, distributing and licensing the GIST system as a research prototype in the scientific community.

By means of further technological development, different classes of systems for dealing with more complex procedural prescriptions can be derived from the GIST results. The drafting of text integrated with pictures – allowing the automatic drafting of user manuals – and the integration of the functional description of components and devices with procedural descriptions – allowing the automatic drafting of manuals out of the formal model of the focused device – are among the envisaged GIST extensions.

Associated partner

UCM (E)

Interest groups

PAB (Provincia Autonoma Bolzano) (I)

INPS (Istituto Nazionale Previdenza Sociale) (I)

Start Date:	December 1993
Duration:	30 months
Resources:	156 person/months
Estimated cost:	1.303.369 ECU

Contact Point

Ms Elisabeth Maier
ITC / IRST
Istituto per la Ricerca Scientifica e
Tecnologica
Localita Pante di Povo
I-38050 Trento
Italy

Tel: +39 461 31 43 45

Fax: +39 461 81 08 51

e-mail: maier@irst.it

Partners

ITC / IRST (Coordinator) (I)

Quinary S.p.A. (I)

UB / ITRI (UK)

OFAI (A)

Methods and Guidelines for Interlinguality in Software Construction (GLOSSASOFT)

Objectives

With five out of the six largest US computer companies now deriving over half their revenue from international sales, leading IT companies have come to recognise the importance of supplying local-language versions of their products. This project aims to produce:

- guidelines on how to structure new software packages and systems
- methods and tools to help adapt existing packages and systems

so that the natural language of interaction can be readily changed.

The project will look at both the software technology and the linguistic aspects of the problem, and will carry out practical localisation case studies, notably of a GUI, an online help system, and a software application. Through this it will produce the guidelines, methods, and tools for software structuring and restructuring, with a view towards broad acceptance and eventual standardisation of the guidelines.

Approach and Methodology

The consortium consists of two research establishments and a university, both of whom have expertise and experience in the area, and three associated systems manufacturers who are interested in the conversion of software products for interlinguality. Further guidance is provided by an Industrial Interest Group, whose members are in frequent contact with the consortium and who voluntarily give feedback on draft deliverables.

The project's approach is novel in that instead of looking at the reuse of linguistic resources within NLP systems, Glossasoft looks at the reuse of other software resources and their integration with linguistic resources. It draws upon current work on software reuse in ESPRIT projects such as 1094 Practitioner, 5327 REBOOT, 5311 BUSINESS, 2487 REDO, the Eureka project ESF ROSE, and other work elsewhere. It reuses existing methods and tools wherever possible.

Changing the language of interaction at first glance may seem trivial, particularly with WIMP interfaces where input is by menu selection and menu headings, prompts, system messages, and similar linguistic software resources are stored in files. Simply replacing the file changes the language of user interaction. But this only works if natural language characteristics are not deeply embedded in the software (as, for example, matching algorithms for word search in a text editor or record look-up in a database system, or assumptions about average word length). For general approaches to prompts and menus there is a need to capture meaning, and use specialised dictionaries of interaction terms to instantiate the interface for a particular language. An essential aspect of the development of interlinguality is usability - at the same time as the language of interaction changed, systems should be more usable. Thus issues of usability will be all pervasive, coming into the case studies, the linguistics studies, and into the tools and methods.

Progress

The project started with an initial phase of two parallel studies taking the foundation technologies of linguistics and NLP, and software structures (architectures and reuse). In conjunction with this a requirements analysis was undertaken using the experience of the industrial partners. This experience was complemented with knowledge gained from the Interest Group which was formed at the project kick-off meeting. Project deliverables for this first phase consisted of the two baseline documents and a commercial requirements document. They were delivered at the end of June 1994.

The second phase addresses the problem of software interlinguality, distinguishing three distinct dimensions to the problem:

- the need to develop guidelines in the two areas for linguistics and software structures
- the need to distinguish the ideal case of developing new software from the pragmatic case of localising existing software
- the need to distinguish the immediate, interface related elements of interaction where linguistic features are obvious, from the deeper components of software like databases where linguistic assumptions are made but may not be easy to disentangle.

This second phase will result in a first version of the guidelines and methods, focussing on the interactive elements in the interface necessary for the localisation of existing software. Project deliverables for this first phase consist of two reports (scheduled for delivery in December 1993) concerning interaction interlinguality for new systems, one containing draft linguistics guidelines, the other draft software structure guidelines. A third deliverable is the draft method for localisation of existing software.

The third phase is the case studies. These will be used to give a first evaluation of the guidelines and methods, and from this experience will come proposals for extensions to the guidelines, methods and tools. These Case studies are thus vital to the success of the project. During the case studies some prototype tools to assist in localisation will be produced.

In the final consolidation phase, the development of the guidelines and methods will be revisited, incorporating the experience gained in the case studies, generalising these to include as wide a class of languages as possible, and adding the deeper features of localisation that will have been revealed in the case studies. A concerted dissemination program will be conducted towards the end of the project, making the guidelines and methods widely available.

Exploitation and Future Prospects

LRE is intended to move European software development towards NLP. Apart from free standing systems or system components like Machine Translation or spelling checkers, there is a pressing need to convert complete systems to work in other languages. This project will provide the enabling technology to do this.

The methods and guidelines to be produced are expected to provide input into the standardisation process, such as activity that follows on the IAP recommendations for internationalisation (ISO 1991). This will be a great benefit to the Community, since all parts of the present Community and eastern Europe could benefit from European information technology with minimal conversion costs. With the conversion procedures that are aimed to be produced, even software not conforming to any eventual standard will be able to be deployed widely within the Community. There will also be a consequential benefit to the Community in export markets. The ability to sell systems in the vernacular is important in some markets, such as Arabic for the Middle East, and will become important in others as these markets expand and open up.

The Interest Group will broaden our perspectives on the requirements for the localisation of software, helping us integrate with other activity in Europe (notably the LISA group), and providing an important channel for dissemination. Results will also be disseminated through presentations at conferences and seminars.

Contact Point

Professor P.A.V. Hall
Department of Computing
Faculty of Mathematics
The Open University
Walton Hall
Milton Keynes MK7 6AA
United Kingdom

Tel: +44 908 652694

Fax: +44 908 653744

Partners

The Open University (coordinator) (UK)

The National Centre for Scientific Research, "Demokritos" (GR)

VTT (FI)

Bull ILO (F)

Associated Partners:

Claris Ireland (IRL)

Hewlett-Packard Hellas (GR)

Start Date:	January 1993
Duration:	26 months
Resources:	111 person-months
Estimated cost:	830.000 ECU

Large-scale Grammars for EC languages (LS-GRAM)

Objectives

Computational grammatical descriptions of natural languages are a key component of any NLP application. In the past much dispersed effort has been spent on developing small to medium-sized grammar fragments but the lack of commonly accepted concepts, methods and tools prevented the creation of widely available and extensible grammatical resources.

LS-GRAM aims to address this problem by developing extensible, well-designed, documented and tested lingware for the nine EC languages based on a common mainstream software platform (ALEP) by re-using linguistic knowledge embedded in existing grammatical descriptions. LS-GRAM is to act as the kernel of an ALEP User Group, providing feedback to the system developers, highlighting formal and computational shortcomings of the system and producing a rule coding manual for lingware developers. The production of lingware and extensive documentation will become part of an ALEP starter kit thus improving the conditions for wide distribution of ALEP.

Approach and Methodology

The project will adopt a staged approach in the execution of the language-specific work. A core consortium of four partners covering three languages (German, English and Spanish) will start in the last quarter of 1993. In a first phase of the project they will do a certain amount of definition work – e.g. with respect to coverage of core grammars, end-of project demonstrator and documentation standards – and methodological work – e.g. concerning the re-use of linguistic (i.e. grammatical, lexical) knowledge embedded in existing large-scale grammars – which will serve as input for the other participants, who will join the project some six to seven months later.

The scope of the project is somewhat larger for the core consortium than for the other participants. For German, English and Spanish coverage will be determined on the basis of corpus analysis, which implies that phenomena which are not normally the focus of theoretical linguists (e.g. dates, parentheticals, appositions) will receive due attention. For the six other languages (Danish, Dutch, French, Greek, Italian and Portuguese) a more limited scope is envisaged; basically the aim here is to develop core grammars, which may serve as boot-strapping material for more ambitious future initiatives.

Furthermore, the project intends to make a contribution to grammar engineering methodology by addressing issues of modularity, extensibility and maintainability. The design of the documentation in view of easing the re-usability of the resources will be a priority issue.

Exploitation and Future Prospects

The creation of well-documented grammatical resources developed in a mainstream formalism covering nine languages will provide an attractive basis for training, research and application-oriented projects to build on. The availability of running grammars is expected to boost the dissemination of the ALEP platform within the NLP scientific community. It is hoped that with this larger-scale multilingual effort the standardisation process with respect to tools and methods in NLP will be reinforced and that it will constitute a milestone in the constitution of an EC-wide NLP infrastructure. Ultimately, the results of this project should encourage industrial product developers to adopt mainstream linguistic descriptions for their NLP applications.

Contact Point

Dr. Paul Schmidt
IAI – Institut zur Fürderung der
Angewandten
Informationsforschung van der
Universität
des Saarlandes
W-66111 Saarbrücken

Tel:	+49 681 39313
Fax:	+49 681 397482
e-mail:	paul@iai.uni-sb.de

Partners

IAI (Coordinator) (D)

University of Essex (UK)

Fundaciun Bosch Gimpera (E)

IMS-CL (D)

Note that the contract signed in late 1993 covers only the core consortium consisting of the four above partners. It is foreseen to extend the present contract in 1994 to allow for coverage of all nine EC-languages.

Start Date	January 1994
Duration	24 months
Resources	105 person-months
Estimated cost	1.574.880 ECU

Multilingual Text Tools and Corpora (MULTEXT)

Objectives

Existing tools for NLP and MT corpus-based R&D are typically embedded in large, non-adaptable systems which are fundamentally incompatible. As a result, there is a serious lack of generally usable software tools to manipulate and analyse text corpora, that are widely available for such R&D tasks as multilingual application development.

The project seeks to address this problem by contributing to the development of such software tools and by creating multilingual text corpora with structural and linguistic markup as software testbeds. In the pursuit of these goals, it will attempt to establish conventions for the encoding of such corpora, as well as guidelines for text software development, building on and contributing to the preliminary recommendations of the relevant international and European standardisation initiatives.

The consortium is committed to make its results, namely corpora, related tools, specifications and accompanying documentation, freely and publicly available. Six major European companies are involved in the project as industrial partners. They will both contribute to the specification and development of the basic tools and provide a first indication of the exploitability of these tools by using them as a basis for building several high-level NLP applications.

Approach and Methodology

At the outset of the project, the consortium will undertake to analyse, test and extend the SGML based recommendations of the Text Encoding Initiative (TEI) on real-size data, and gradually develop encoding conventions specifically suited to multi-lingual corpora and the needs of NLP and MT corpus-based research. To manipulate large quantities of such texts, the partners will develop conventions for tool construction and use them to build a range of highly language-independent, atomic and extensible software tools.

These specifications will be the basis for the development of two major software resources, namely (a) tools for the linguistic annotation of texts (e.g. segmenters, morphological analysers, part of speech disambiguators, aligners, prosody taggers and post-editing tools), and (b) tools for the exploitation of annotated texts (e.g. tools for indexing, search and retrieval, statistics). This software will be implemented under UNIX, while its specific properties should facilitate portability to other systems. Moreover, it will be integrated by means of a common user interface into a text corpus manipulation system expected to provide the basic functionality needed in academic or industrial corpus research. For the overall software design as well as the development of specific components. MULTEXT will capitalise on the preliminary results achieved in the ALEP project.

By using the emerging software tools, the consortium plans to produce a substantial multilingual corpus, including parallel texts and spoken data, in six EC languages (English, French, Spanish, German, Italian and Dutch). The entire corpus will be marked for gross logical and structural features; a subset of the corpus will be marked and hand-validated for sentence and sub-sentence features, part of speech, alignment of parallel texts, and speech prosody. All markup will have to comply with the TEI-based corpus encoding conventions established within the project. The corpus will also serve as a testbed for the project tools and a resource for future tool development and evaluation.

An application programming interface will facilitate the coupling of the progressively refined software and data components with several existing language application systems or prototypes. In particular, the industrial partners plan to develop extraction software for lexical and terminological information to complement and improve their Terminology Management, Information Retrieval or Machine Translation systems. Some effort will also be devoted to a prototypical application for testing and comparing successive versions of a Machine Translation system.

Exploitation and Future Prospects

Text-oriented methods and software tools have come to be of primary interest to the NLP community. It is therefore expected that the availability of basic multi-lingual tools and data will improve and extend R&D across a wide range of disciplines, including not only the various areas of NLP (language understanding and generation, translation, etc.), but also fields such as speech technology, language learning, lexicography and lexicology, literary and linguistic computing, information retrieval, etc. By feeding the results into several commercial application systems/prototypes, the project is expected to show the potential of state-of-the-art methods in corpus linguistics for improving industrially relevant language systems and services.

By interacting with prominent research organisations and initiatives inside and outside the EC, it is hoped that MULTEXT's approach will receive the attention of a wide international forum. In a longer term perspective, it can be anticipated that this project will strengthen the methodological and technological foundations for the uniform representation, annotation and exploitation of textual information.

Contact Point

Mr Jean Veronis
Laboratoire Parole et Langage
U.R.A. 261 CNRS
Universitè de Provence
29, Avenue Robert Schuman
F-13621 Aix-en-Provence Cedex 1

Tel: +33 42 20 43 56 (secretary)

Fax: +33 42 20 59 05

e-mail: veronis@grtc.cnrs-mrs.fr

Partners

CNRS, Paris (Coordinator) (F)

EUROLANG-SITE, Maisons-Alfort (F)

Siemens Nixdorf-CDS, Barcelona (E)

Digital Equipment, Nieuwegein (NL)

CAP debis Systemhaus KSP, Munich (D)

University of Pisa (ILC/CNR) (I)

University of Edinburgh (HCRC/LTG) (UK)

ISSCO, Geneva (CH)

Associated Partners

Siemens Nixdorf IS, Munich (D)

Universitaet Muenster (D)

Rank Xerox Research Center, Grenoble (F)

Universitat Autonoma de Barcelona (E)

Universitat Central de Barcelona (FBG) (E)

Universiteit Utrecht (NL)

Start Date:	January 1994
Duration:	26 months
Resources:	238,5 person-months
Estimated cost:	3.210.000 ECU

61-004

Multi-Language Pronunciation Dictionary of Proper Names and Place Names (ONOMASTICA)

Objectives

The availability of large pronunciation dictionaries is becoming an important factor for the development of many applications in speech technology. Future automated operator services, telephone banking and map guidance systems, for instance, will depend heavily on the capability of producing correct pronunciations for names of various categories.

The objective of the project is therefore to make available quality controlled pronunciation lexicons in machine readable form (CD-ROM), for a total of nine EC languages, namely Danish, Dutch, English, French, German, Greek, Italian, Portuguese, and Spanish, as well as Norwegian and Swedish. These lexicons are intended to include city and town names, street names, family names, company names, product names, up to a total of 1.000.000 names per language. Besides lexical data, the results of the project will also include optimised sets of letter-to-sound rules developed in the project as an accelerator to human editing.

The final version of the ONOMASTICA lexicon will contain transcriptions coded as unique IPA numerical references, thus providing simple, comprehensible transcriptions which allow native and non-native speakers to produce adequate and natural pronunciations of names. Furthermore, the transcriptions should be usable as input for speech synthesis systems and/or lexicons for speech recognition applications.

Approach and Methodology

The approach in the project is to define a dictionary consisting of names and their phonetic representation. To create this dictionary, lexicographers will initially select from names lists provided by the associated partners, 20.000 to 50.000 names for the language. The dictionary would then be generated directly by hand, by an expert phonetician transcribing the conventional pronunciations of the names. On the basis of this initial sample, it will be possible to write sets of grapheme-to-phoneme conversion rules, and evaluate these rules in preparation for (semi)-automation of the further lexicographic work. An alternative pre-processor based on neural computing methods will also be investigated.

Automatically generated transcriptions will receive the quality label Band III as transcriptions not yet hand-checked. In subsequent stages of the project, these will be checked and edited where necessary and promoted to Bands I and II, defined as transcriptions judged to be correct to the best of a competent phonetician's knowledge, or judged to be acceptable to a native speaker/listener. The level of transcription adopted by the consortium is a broad phonetic level, centred on the marking of important allophonic contrasts, lexical stress on multi-syllabic names, and syllabification and word boundaries.

The project, in its later stages, will also seek to investigate the problems of exchanging national names amongst the partners to create a matrix of 'native-ised' pronunciations for each (thereby) foreign name in each other language.

Exploitation and Future Prospects

The primary deliverables from this project will be a documented set of multi-language, machine readable CD-ROM pronunciation dictionaries for European city and town names, street names, family names, company names and product names, including the rules that have been employed to produce them.

These results will constitute a valuable linguistic resource which allows language products to handle names correctly. Benefits will be felt in applications such as automated directory inquiry systems which can provide telephone numbers using advanced machine dialogues, recognising the desired name and address; map information and guidance systems which can recognise and synthesise names accurately; and systems such as talking newspapers and books (for the blind) which can accommodate occurrences of names without pronunciation errors.

The project is operated by a group of eleven academic partners with eleven associated partners from the European telecommunications industry. In themselves this group represents a major user group for potential downstream exploitation of the results of the project.

Progress

The rules have been tuned to the names of each language under consideration, and are now undergoing efficiency tests. Based on the results of the latter, they will be checked against the self-learning code developed in parallel. The first out of three lexicon compilation phases is reaching its end with the provision of approximately 250.000 entries per language. The consortium has held research seminars, for instance on onomastics, and demonstrated an application of ONOMASTICA lexicon data at EUROSPEECH '93.

Contact Point

Prof. Mervyn A. Jack
University of Edinburgh, CSTR
South Bridge
Edinburgh EH1 1HN
United Kingdom

Fax: +44 31 226 2730

Tel: +44 31 650 2783

Partners

CSTR, University of Edinburgh (coordinator) (UK)

Speech Technology Centre, Aalborg University (DK)

Ecole Nationale Supèrieure des Tèlècommunications - ARECOM, Paris (F)

Institut für Fernmeldetechnik, Technical University of Berlin (D)

Dept. of Electrical Engineering, University of Patras (GR)

Istituto di Linguistica Computazionale del CNR, University of Pisa (I)

Catholic University of Nijmegen (NL)

Instituto Nacional de Engenharia de Sistemas e Computadores, Lisbon (P)

Univ. Politècnica de Madrid (E)

SINTEF DELAB, Trondheim (NW)

Kungl Tekniska Högskolan, Stockholm (SW)

Associated Partners

BT Laboratories, Martlesham (UK)

Jydsk Telefon, Aarhus (DK)

France Telecom (CNET), Lannion (F)

Deutsche Bundespost Telekom, Darmstadt (D)

Intrasoft, Athens (GR)

CSELT, Turin (I)

PTT Research, Leidschendam (NL)

Telefones de Lisboa e Porto, Lisbon (P)

Telefónica, Madrid (E)

Norwegian Telecom Research, Kjeller (NW)

Telia (Infovox), Solna (SW)

Start Date:	January 1993
Duration:	24 months
Resources:	302 person-months
Estimated cost:	3.577.280 ECU

European Network of Repositories for Linguistic Resources (RELATOR)

Objectives

The language industries of the future will rely heavily on the availability of large scale language resources e.g. corpora, speech databases, dictionaries, linguistic descriptions – together with appropriate standards and methodologies. Ready access to harmonised databases of language data and rules would not only provide a direct benefit to research and development efforts across a wide range of private and public organisations, but would also foster fruitful academic and industrial co-operation.

The project aims to define a broad organisational framework for the creation, collection, verification, normalisation and re-distribution of the language resources for both written and spoken language engineering (LRs in short) which are necessary for the development of an adequate language technology and industry in Europe, and to determine the feasibility of creating a co-ordinated European network of repositories which would perform the function of storing, disseminating and maintaining such resources. This activity is intended to contribute towards the long term goal of making large scale LRs widely available to European organisations involved in R&D and educational activities.

Approach and Methodology

The overall approach and the results which the project intends to achieve can be summarised as follows:

- to create structured, publicly available catalogues of existing linguistic resources, using and extending the information already collected by various international and national survey initiatives;
- to evaluate the present European situation, comparing what is available with the most urgent needs of the European R&D and teaching communities, and then to formulate recommendations for a concerted European action in the field of reusable resources for natural language and speech;
- to discuss with the relevant actors (e.g. owners of resources, producers, private and public users, funding bodies, scientific and professional associations) the various aspects of the problem, their needs and requirements, the possible solutions, their willingness to co-operate, and the conditions for a joint European action;
- to identify, describe and evaluate at various levels (e.g. organisational, technical, legal) alternative methods and structures which could ensure the establishment, management and maintenance of a European repository of reusable LRs, and their dissemination to the various types of users;
- to experiment with the collection and dissemination of existing LRs using (i) a distributed electronic network and (ii) CD-ROM pressing facilities, with the aim of encouraging the reuse of already

available resources, and also of acquiring experience which will feed into the formulation of final recommendations;

- to present final recommendations for establishing a collaborative infrastructure that will act as a collection, verification, management and dissemination centre, built on the foundation provided by existing European structures and organisations.

Assessing Existing Resources: carrying out a review of what LRs currently exist, both in Europe an elsewhere. The goal of this survey is not to produce a comprehensive, exhaustive catalogue of such resources, but rather to assess which needs of the various European languages are still not satisfied by the available resources, and to compare and characterise the situations of the different languages. The results of this evaluation effort will provide the basis for the general recommendations (see below).

Needs Analysis: determining the main resource needs of European actors involved in RTD training and system development; discussing the various aspects (e.g. legal, financial, organisational problems; participation and role of different types of public and private actors) of the actions required to meet the needs for LRs in Europe, as a basis for defining an overall organisational framework for the development of adequate LRs in Europe.

Experimental Implementation: testing the usefulness and feasibility of a distributed resource repository by implementing an infrastructure on which will be mounted a set of LRs; in particular we will experiment with the dissemination of LRs using ELSNET's existing infrastructure for LRs: (i) a wide-area network running the AFS server software, and (ii) the formatting, mastering and distributing of data by CD-ROM.

Recommendations: making detailed recommendations for the creation, management, and maintenance of a distributed, managed repository of reusable LRs, based on a detailed analysis and evaluation of the alternatives.

Exploitation and Future Prospects

The goal of the project is the co-ordinated collection and distribution of LRs, promoting awareness of the need for creating widely available LRs, and the promotion of consensus on an overall European strategy. Consequently, dissemination activities are central to the project. The project consortium comprises representatives of major European-wide bodies and associations, most notably ELSNET, ESCA and EACL, and will be assisted by an industrial steering committee composed of representatives of leading IT companies, publishers, PTTs and other providers of electronic information services.

The action will be carried out in co-operation with relevant European groups and with on-going initiatives such as EAGLES and EURO-COCOSDA, and will imply amongst other things an analysis of existing international structures. It is expected that the experimental activities carried out within the project and the recommendations for further larger-scale operations will contribute to the establishment of a collaborative language infrastructure covering all Community languages.

Contact Point

Prof A. Zampolli
Dipartimento di Linguistica
Computazionale
University of Pisa
Via della Faggiola 32
I-56100 Pisa
Italy

Tel:	+39 50 56 04 81
Fax:	+39 50 58 90 55
e-mail:	eagles@icnucevm.cnuce.cnr.it

Partners

University of Pisa (I)

LIMSI-CNRS, Paris (FR)

University of Edinburgh (UK)

DFKI - Univ. of Saarland, Saarbrücken (D)

Associated Partners

CST Copenhagen (DK)

Institut de la Communication Parlee,
Grenoble (F)

Start Date:	January 1994
Duration:	18 months
Resources:	55,5 person-months
Estimated cost:	592.000 ECU

Reduction of Noise and Silence in Full Text Retrieval Systems for Legal Texts (RENOS)

Objectives

The problems of accessing texts in a large textual database are not restricted to the legal world, although they are as acute here as in most other spheres of activity. Existing retrieval mechanisms depend upon the user being able to formulate his query in the form of strings of keywords included in an inverted list (a primitive index), which restricts usage of the texts as it ignores some properties of the legal sublanguage.

The RENOS project aims to develop software modules capable of being integrated into existing Full Text Retrieval Systems (FTRS) which will reduce the levels of "noise" and "silence" of such systems when applied to legal texts. "Noise" is defined as the retrieval of texts of little or no relevance to user queries, while "silence" is defined as failing to retrieve relevant texts from the database. The software modules will implement a semi-automatic methodology for identifying legal terms (single-word and compound terms) in legal texts originating from several European member states by statistical means and by morphological and linguistic analysis.

Approach and Methodology

The approach adopted in the project is the creation of an "intelligent inverted list", which comprises a lexicon of single-word and compound terms, a hierarchically arranged conceptual network and a constituent grammar. Lexicon entries will be linked to nodes in the network and these nodes – "concepts" – will form the basis of text retrieval. Constituent grammars will offer linguistic criteria for identification of compound terms and ambiguous terms, i.e. words used both as a legal term and in the general language meaning.

The lexicon will contain a framed representation of single-word and compound legal terms, which will be stored by their stems together with pointers to inflectional patterns. Nodes in the conceptual network will consist of semantic classes pertaining to legal terms – "concepts" – organised in a tree structure. Pointers from lexical entries to the concepts in the network will be established, synonymous terms pointing to the same node. The constituent grammar will contain rules for the identification of compound terms and disambiguation of the meaning (legal or general) of single word terms in context.

The components of the network will be manually built in the prototype system, following automatic extraction of an initial set of terms from a corpus of legal text containing legislation common to Community countries. Part of the software to be built will establish the links between network concepts (nodes) and the corpus by applying grammar rules on appropriate corpus segments. Another part will implement a mini Text Retrieval System using the

Intelligent Inverted List, demonstrating its benefits over traditional methods of text retrieval. Evaluation stages will quantify the performance of the RENOS system with respect to existing FTRSs.

Exploitation and Future Prospects

The end result of the RENOS project will be a piece of software which, with some additional development work, may support a multilingual FTRS, and the two private companies in the consortium, both legal information providers, plan to exploit this directly. Databank S. A. will explore the possibilities of incorporating tools and methodologies in the NOMOS database, and SOGEI will similarly attempt to integrate the conceptual legal term network into some of its existing products and services.

The collection of legal terms in three European languages is a key feature of the project, together with the evaluation and refinement of automated tools for the acquisition of terminological resources by statistical means. Extension to other languages and subject areas (engineering standards, medical texts) is envisaged.

Incorporation of the intelligent inverted list demonstrated in RENOS into existing FTRSs will greatly improve their query mechanisms, and the RENOS system could eventually be commercialised via direct sales to text retrieval companies and information providers.

Contact Point

Dr Leonidas Bardis
Databank S.A.
124, Kifissias Ave. & Iatridou St.
11526 Athens
Greece

Tel: +30 1 649 4830

Fax: +30 1 649 0012

email: lbardi@nrcps.ariadne-t.gr

Partners

Databank S.A (Coordinator) (GR)

Intrasoft S.A. (GR)

SOGEI (I)

Instituto di Linguistica Computazionale (I)

CEF Management Research Centre (DK)

Institute for Language and Speech Processing (ILSP) (GR)

Start Date:	December 1993
Duration:	20 months
Resources:	116 person-months
Estimated cost:	1.075.000 ECU

The Reusability of Grammatical Resources (RGR)

Objectives

The lack of appropriate tools for language engineering constitutes a serious impediment to the wide-scale commercial exploitation of NL research for product development. It also limits the value of research systems that could be used for developing, testing and improving linguistic theories. To overcome this problem, appropriate formal concepts and a representation language for the abstract specification of grammatical knowledge must be developed.

This project aims to provide extensions to standard unification grammar formalisms on the basis of linguistic and computational considerations. The extensions can be thought of as consisting of a library of datatypes. A datatype is a type of object that is used to represent information or knowledge combined with the operations on that type of object to manipulate that information. The project will deal both with the syntax that is employed to refer to objects and operations, and with the computational implementation of these objects and operations.

Approach and Methodology

Beginning with an inventory of the datatypes that best serve linguistic descriptions in mainstream frameworks such as HPSG, GB, LFG and CG, the project will produce a shortlist of the most prominent datatypes to be included in the library, and will provide specifications to implement the selected datatypes within the core of an implementation of the formalism of the ALEP platform, or as extensions to that formalism. The specifications will form the basis for a version 0 implementation of the datatype library. In the final phase, the library will be assessed, re-implemented and made available for general distribution, together with the accompanying documentation.

Exploitation and Future Prospects

The project will result in a well-documented library of datatypes to accommodate linguistic descriptions. The library will facilitate the writing of large-scale grammatical descriptions with a realistic coverage. It will improve the transparent representation of linguistic knowledge, thus increasing the possibilities of (re)using grammatical descriptions in various applications. Finally, the library will also reduce the costs of developing large scale grammatical descriptions.

The results of the project are expected to be of interest to companies that engage in development of grammatical resources for commercial or research purposes. By expanding the datatypes to include those used in the major linguistic frameworks, the project will open the computational field to mainstream linguistics research. In addition, it will make the ALEP formalism and its implementations an attractive option to both the scientific and the commercial computational linguistic community.

The project aims at a widespread acceptance of the library and the related ALEP formalism. To achieve this,

- the library will be accompanied by extensive documentation;
- a number of promotion activities will be undertaken. These comprise the presentation of project results at national and international conferences and a workshop to familiarise potential users with the use of the datatype library;
- the prototype library and the scientific reports will be made freely available.

Progress

Work on the general requirements specification for a library of datatypes has been completed. The project has delivered a set of reports comprising: an overview of datatypes in HPSG, GB, LFG and CG; selection criteria; a description of the application of the criteria to the candidate datatypes; an ordered list of suitable datatypes selected for treatment in this project. Work is under way on the formal specification for the selected datatypes, that will form the basis for the implementation of the datatypes. A report on this formal specification is to be delivered by the end of 1993.

Contact Point

Herbert Ruessink
Stichting Taaltechnologie
Trans 10
NL-3512 JK Utrecht
The Netherlands

Tel: +31 30 536 369

Fax: +31 30 536 000

Partners

Stichting Taaltechnologie, Utrecht (coordinator) (NL)

HCRC, University of Edinburgh (UK)

Universität des Saarlandes, Saarbrücken (D)

Instituut voor Taal- en Kennistechnologie, Tilburg (NL)

Start Date:	January 1993
Duration:	25 months
Resources:	106,5 person-months
Estimated cost:	790.000 ECU

A Simplified English Grammar and Style Checker/Corrector (SECC)

Objectives

Many companies are nowadays operating in an international, multilingual environment. In such an environment, fast and effective communication in the language of the customer is a key factor for success. Using a reduced subset of a language, companies are able to produce consistent, unambiguous and easily understandable handbooks, technical documentation and training material. The task of technical writers taking advantage of simplified language would be greatly simplified if they could rely on specific writing tools for simplified languages. The SECC project intends to develop precisely such a tool, namely a grammar and style checker for simplified/controlled English (SE).

The SECC tool is meant to serve two purposes. In the first place, SECC will be a writing tool for technical writers who have to produce easily readable, unambiguous texts (manuals, for instance) that should be understandable by a wide audience of non-native speakers/writers of English. In the second place, the tool should be able to serve as a front-end or pre-translator for machine translation products helping to improve translation quality and reduce post-editing work by simplifying the input.

The tool will perform syntactic and lexical (terminological) checking, on all levels of the text. As such, it goes beyond the usual upper boundary of the sentence: the syntax (layout) of paragraphs, sections and overall text will also be checked against the SE rules governing those levels. Special attention will be paid to mistakes by non-native (viz. Dutch, French and German) writers of SE.

The different interfaces for the user will form an important subpart of the project. The SECC tool will run both in batch mode (checking of completed texts) and in interactive mode (checking of subparts of a text while it is written), from within the Interleaf5<SGML> Desktop Publishing package on Sun workstations.

Beside the major objective of developing the tool, the project will also set up an industrial interest group working together on international developments related to controlled language, and organise an international SE workshop, bringing together developers and users to propagate the use of SE.

Approach and Methodology

The SECC tool will be based on existing NLP technologies, being built within an existing machine translation framework, and it will reuse NLP and linguistic resources. The task of the tool will be to translate from English to a subset of it (SE); in this respect, SECC will not limit itself to the output of diagnoses of mistakes, it will also attempt to correct (translate) erroneous sentences as much as possible. The tool will reuse the analysis component for English (grammar and lexicon) of the Metal MT system in order to do a thorough syntactic analysis of the input. For the SE rules and lexicon, again existing sources will be reused. A solid 140-rule grammar of SE developed in the context of the telecommunication subdomain of telephony, as well as a union of electronically available existing basic SE lexicons plus technical terminology, will together form the "transfer" modules of the checker/corrector. For generation, SECC will reuse the English synthesis module of Metal. This MT approach to syntax checking has already been successfully applied (albeit only experimentally) to German in the context of the ESPRIT TWB project, using the same system.

Interface developments will include the complexities of communication between the DTP package and the NLP application, user-friendly interfaces using the Motif standard, hypertext-like presentation of the checker's output, and internal representation of the output using the SGML standard. SGML-related tools will also be used to develop the checking and correcting modules beyond the sentence level.

Exploitation and Future Prospects

The SE Grammar and Style Checker/Corrector will be a powerful tool that can be used by any organisation with strong needs for efficient communication (text production and translation) in an international context or market.

First of all, the tool will be used in house by one of the partners in the production process of texts for technical courses and telecom user documentation. In addition, it is planned to offer SECC as a product all over Europe through the commercial divisions of the partners involved.

In order to get as broad a dissemination of results as possible, descriptive results (user requirements, overall system approach, academically interesting results relating to grammar, lexicon, restricted languages, etc.) will be made public.

As to the future technological prospects, SECC will be part of the EUROLANG developments, aiming at offering a widespread European NLP platform, based on the same technology as SECC.

Contact Point

Mr. Geert Adriaens

Siemens Nixdorf Software Center LiËge (CSL)

Rue des Fories 2

B-4020 LIEGE

Belgium

Tel: +32 41 20 17 07

Fax: +32 41 20 16 42

e-mail: gad@csl.sni.be

Partners

SNI CSL (Coordinator) (B)

Cap Gemini Innovation France (F)

Alcatel Bell (B)

University of Leuven (KUL CCL) (B)

Associated Partners

Sietec Systemtechnik Munich (D)

Start Date:	December 1993
Duration:	30 months
Resources:	125 person-months
Estimated cost:	1.560.952 ECU

Selecting Information From Text (SIFT)

Objectives

The problem of access to textual information is one faced by every organisation. SIFT will demonstrate a technology which can directly address this task.

The project aims to construct a demonstration intelligent help system for online computer software manuals based on two key ideas: the Vector Space Model of information retrieval on the one hand and the use of distributed patterns to capture the meaning of textual information on the other. The final prototype will accept a user's query in natural language concerning the software and return a list of pointers into the manual texts indicating where passages answering the query might be found. These will be arranged in descending order of relevance to the query, allowing the user to investigate the most promising parts of the text first.

The project will also serve to demonstrate the usefulness of distributed patterns in practical Natural Language Processing systems and their compatibility with existing work on lexical databases and robust lexicalistic parsing.

The combination of VSM retrieval mechanisms and distributed patterns has already been demonstrated in a working retrieval system. However, this was created largely by manual means. SIFT will provide the technology for generating such a system automatically

Approach and Methodology

The SIFT system will consist of two main components. The document processing component will analyse an SGML tagged computer manual and associate with its different sections, subsections and individual sentences distributed patterns capturing the meaning – in gist – of those textual units. The interactive query processing component will accept a user's input query and produce as output an ordered list of pointers to text portions.

The key ideas in the project are the use of robust lexicalistic parsing (of manual texts and user queries), the assignment of semantic cases to syntactic constituents and the extraction of distributed representations from machine readable dictionaries.

Exploitation and Future Prospects

The techniques to be used in SIFT are intended to be applicable to a wide range of related text processing tasks and could be incorporated into other products such as stylistic checkers, summarisation engines and machine assisted translation tools.

It is also expected that SIFT will yield theoretical insight into the applicability of automated processes to natural language processing, e.g. how distributed representations can be used to capture both word and sentence meanings, how a large, robust and general purpose semantic lexicon can be constructed automatically, and whether robust lexicalistic partial parsing is possible.

Plans for the commercial exploitation of SIFT technology in a commercial product are being investigated.

Contact Point

Dr Richard Sutcliffe
Dept. of Computer Science
and Information Systems
University of Limerick
Limerick
Ireland

Tel: +353 61 333644 ext. 5006

Fax: +353 61 330876

email: sutcliffer@ul.ie

Partners

University of Limerick (coordinator) (IRL)

University of Amsterdam (NL)

University of Heidelberg (D)

Associated Partners

Lotus Development Ireland (IRL)

TecnoLingua S.L. (E)

Start Date:	December 1993
Duration:	24 months
Resources:	123 person-months
Estimated cost:	1.177.000 ECU

Semi-automatic Indexing System for Technical Abstracts (SISTA)

Objectives

The SISTA project addresses the area of Document Abstracting and Indexing. SISTA is a two year project to develop an NLP-based framework to assist in automatic indexing of technical abstracts written in English.

The project's major deliverable will be a PC-implemented prototype that automatically reads a technical abstract, isolates and prioritises its main index terms and, where possible, marks the location of the terms within the abstract. An interactive tool then allows the indexer to confirm or edit the terms and their locations. The overall indexing system will therefore be semi-automatic and will represent a realistic compromise between productivity and accuracy.

Drawing on the consortium's expertise in abstract publishing, NLP research and advanced PC software development, SISTA's central objective will be to:

- research, design and build a suite of prototype PC-based tools to index, semi-automatically, technical abstracts written in English

The practical purpose of the above tools is to:

- improve the productivity and consistency of the indexing process
- improve the quality of index users' facilities
- contribute to the " equality of access" ideal for databases, especially text databases in machine-readable form, covering arbitrary subject matter in unrestricted discourse
- provide an NLP-based framework for more diverse applications such as on-line documentation and technology-based training

Although SISTA focuses on English as its application language, a major concern will be to develop a solution that will allow the methodology to transfer directly to other languages.

Approach and Methodology

SISTA's indexing problem is provided by a user group of European secondary service publishers who believe that immediate improvements in productivity and consistency can be achieved by developing current NLP technology in a carefully targeted way. Being user-led, the consortium has set itself commercially significant but technologically feasible goals whose impact can be assessed objectively.

SISTA's NLP technology represents a novel combination of classical symbolic processing and statistical analysis. Firstly, each pre-indexed abstract undergoes a surface-oriented parse which reveals basic sentence structure. Next, potentially diagnostic constituents of the sentence structure are isolated. Then, by comparing these diagnostic constituents with the correct index entries, across the corpus of pre-indexed abstracts, a statistical model is developed. Lastly, the model's robustness is tuned by a global text matching algorithm. The final system will therefore take an abstract as input and return a prioritised list of index terms. The location of the terms within the abstract is then marked, using the ISO standard SGML notation, according to linguistic data provided by the parser.

The technological challenge is to provide, for each abstract, a prioritised list of index terms, drawn from a thesaurus. The strategy is to use a large corpus of pre-indexed abstracts to " train" a statistical model. This model will then be tested against a second set of pre-indexed material. The refined model will then be given a prototype PC-implementation which will be evaluated on-site by members of the user group.

Clearly SISTA's technological strategy presupposes experience in both symbolic processing and stochastic methods, not only at the level of academic research, but also in building commercial implementations. These key skills are represented strongly both at the research and the technology transfer level.

Progress

The SISTA project has concentrated so far on building increasingly sophisticated models, achieving gradually better results. Initially, a " base model" was built which used all the words occurring in an abstract to predict one or more of 90 categories. Subsequently, a more advanced model has been built to suggest key term assignments from a possible 600 key terms. This model uses text tagged with parts of speech and parsed to identify simple noun groups. This gave a precision of 85% (precision is the proportion of the suggested key terms which are correct). Further work on the model will improve recall (the proportion of the correct key terms which were actually suggested) and evaluate different linguistic analyses.

Meanwhile, the indexers' interface has been developed to run under MS-Windows on the PC which will be beta tested by the members of the user group before the end of 1993.

Exploitation and Future Prospects

Although SISTA focuses on a specific well-defined commercial problem, the proposed solution has widespread generic application beyond publishing to such diverse sectors as technology-based training and on-line documentation. The beneficiaries of a semi-automatic indexing tool are index developers and index users. Improvements in productivity will bring direct resource savings to index developers. In addition, it is one of SISTA's aims to evaluate any improvements in indexing consistency and quality. Improved consistency will bring direct savings by reduced user time and indirect benefits of increased ease of use to end users.

The most direct qualitative result will be the development of improved text retrieval methods by importing computational linguistics techniques into current information retrieval technology. This process will also inform and benefit current computational linguistics research, since it will allow the validation of tools in real-life text processing applications.

It is hoped that SISTA will contribute to the process of change in computational linguistics by demonstrating that transfer of NLP technology can yield linguistically well-motivated yet robust systems. In particular, SISTA's twinning of NLP technology with the text representation formalism of SGML may stimulate collaboration between these historically separate research communities. It is hoped that this stimulus will be felt by researchers working on European languages other than English.

Although for the purposes of the project SISTA's subject domain is technical abstracts, the project's techniques will apply directly to abstracts dealing with other topics. In addition, SISTA's results will prove useful to other quite different applications and growing markets such as on-line documentation and technology-based training, both in English and other languages.

Contact Point

Mr Phillip Joyce
Brain Training
The Jeffreys Building
St John's Innovation Park
Cambridge CB4 4WS
UK

Tel:	+44 223 421 823
Fax:	+44 223 423 404

Partners

Brain Training (coordinator) (UK)

HCRC, Univ. of Edinburgh (UK)

AIC Ltd (IRL)

Elsevier Science Publishers B.V. (subcontractor) (NL)

Start Date:	January 1993
Duration:	26 months
Resources:	65 person-months
Estimated cost:	590.000 ECU

62-058

Speech Recognizer Quality Assessement for Linguistic Engineering (SQALE)

Objectives

Evaluation is a key component of every technology, in that it allows to assess the performance of systems using a given technology and to match user requirements against system performance. In the field of linguistic engineering – encompassing both language and speech technology – it is a most urgent task to elaborate and establish common and widely accepted evaluation methods.

The SQALE project intends to contribute precisely to this effort; in fact, it will develop an assessement paradigm for large vocabulary, speaker independent, continuous speech recognition in Europe, taking into account the distinctive characteristics of a multilingual environment and identifying the problems it raises. Also, the project will begin the definition of guidelines for future assessement actions. If the SQALE project proves successful, these guidelines could be extended to an evaluation paradigm for future large scale European language/ speech programs.

The project will also directly contribute to the assessement and evaluation of NLP systems in at least three ways: a general framework will be established for comparing machine generated output with reference corpora; a first step will be taken toward handling real-word phenomena, such as false starts and hesitations; the effects of differing test set perplexities across various European languages will be quantified. The SQALE experiment will therefore not only extend European standards in speech recognition assessment (which are limited to isolated and connected word systems, without a direct link with language models) but also initiate the necessary and much awaited integration between speech and NL assessment methodologies.

Approach and Methods

The project takes into account the experience gained by the partners in the 1992 DARPA RM and WSJ evaluations in order to investigate how the US protocols can be improved and extended into a multilingual experimental design, as required for a European approach.

The basic idea of the project is to form a small consortium, made up of a coordinating laboratory – having a high technical expertise in the field – and three other laboratories testing their "in house" recognition systems. The "testing" laboratories are located in three different countries, where three different languages are spoken. Hence two dimensions of the research paradigm are investigated: the recognition algorithms (at least three) and the languages (at least three). In particular, the experiment will focus on two independent research questions:

- the merits of different recognition algorithms applied to the same data, and

- the relative difficulties in speech recognition across different languages.

Having multiple sites applying their algorithms on the same database makes it possible to discuss the merits of different methods on the same data. Testing the same algorithm on different databases in different languages will reveal the relative difficulties of speech recognition for different languages, and the degree of robustness of the algorithm with respect to a given language.

Each testing laboratory will be responsible for providing data in its own language – both written and spoken corpora – for assessing its systems according to a commonly accepted protocol and for performing the assessement procedure for English and at least one other language. The coordinator will organize the assessement experiment and will be responsible for timely distribution of training and test materials, and for gathering of the tests results. He will also score the recognizers output and analyze the results.

The high quality of the three test sites and their recognition systems (all three labs proved to perform at the top level in the DARPA 92 bench mark test) and the high technical standard of the coordinating laboratory are considered essential ingredients for the success of the project.

Exploitation and Future Prospects

SQALE intends to bridge the gap between the state-of-the-art in commercial systems assessment – as examplified by the SAM Esprit project – and the state-of-the-art in research systems assessement – as represented by ARPA. It will therefore have a direct relevance to current leading edge research and development, and should also have a pull through effect on future application-driven and technology-driven research. Furthermore, SQALE will operate in a multilingual European context and will therefore go beyond the current ARPA scope. Cross-language assessment and evaluation have never been performed on this scale previously: SQALE will be a pioneer project in this respect.

As far as more immediate and practical results of the project are concerned, the dissemination of the following material is envisaged (primarily through EAGLES):

- the speech corpora, including the speech signal and the associated transcription, the lexica and the text corpora;
- the results obtained by each testing participant in its own language and in the common language;
- the guidelines and recommendations on how to conduct and organize systems evaluation in a multinational, multilingual context.

These results will constitute a baseline from which it will be possible to improve the methodology, enlarge the number of participants, augment the difficulty of the tasks and ensure the coordination with closely related research areas, such as written language processing and machine translation.

A primary basis for interaction between speech and NL systems will be represented in fact in the near future by the common use of text corpora and statistically based language models. The development of common assessement methodologies and protocols will be equally relevant for NL and speech integration.

Contact Point

Dr. H.J.M. Steeneken
TNO Institute for Human Factors
Kampweg 5
P.O. Box 23
3769ZG Soesterberg
The Netherlands

Tel: +31 3463 56269

Fax: +31 3463 53977

e-mail: hjms@izf.tno.nl

Partners

TNO – Institute for Human Factors (Coordinator) (NL)

LIMSI – CNRS (F)

Philips – Aachen (D)

CUED (Cambridge University Enginering Department) (UK)

Start date:	December 1993
Duration:	18months
Resources:	49 person-months
Estimated cost:	435.217 ECU

62-070

A Testbed Study of Evaluation Methodologies: Authoring Aids (TEMAA)

Objectives

Although the first attempts to evaluate specific natural language processing (NLP) systems date back to as early as the sixties, no satisfactory, comprehensive evaluation method has been developed over the last three decades. The absence of generally accepted quality criteria and benchmarks for NLP systems has led to a situation where system developers, researchers, sponsors and users of NLP products and services are forced either to develop their own set of evaluation criteria and techniques on a case by case basis, or to rely on evaluation methods that were developed for a particular NLP system or component and are therefore not easily exportable.

This project, which will be carried out in close collaboration with manufacturers of language engineering products, intends to remedy this situation by working towards a general framework for an evaluation methodology for NL products and projects. The proposed methodology will be used to set up a concrete evaluation package for spelling checkers. The project will also develop a preliminary set of evaluation criteria and methods for grammar checkers, taking into account products as well as ongoing development and research projects in the area. Finally, ongoing projects on information extraction and retrieval will be followed, to get feedback from this area to the development of the general framework.

Approach and Methodology

Preliminary work involves discussions with customers to establish what for them are important factors in judging the acceptability of the products in question. On that basis, and in the light of previous work in the context of the EAGLES Evaluation and Assessment group, the project will define a list of "evaluanda", and of associated criteria determining acceptability.

In parallel, the project will examine the evaluation methods already used by manufacturers and carry out a survey of existing, documented methods with a view to defining tests designed to collect evidence relevant to each of the criteria identified. The information gathered will be critically examined in the light both of previous work carried out by the EAGLES group and of insights gained from quality assessment and control work in industry in general.

The critical examination will lead to a preliminary definition of a set of evaluation methods for spelling checkers. These evaluation methods will be validated in an experimentation cycle, which will lead to modification of the methods proposed and new experimentation. During the experimentation phase, several European languages will be taken into account. Care will be taken to ensure that the tests defined are independent of any specific software support.

For the second application area, grammar checkers, a partial evaluation methodology will be developed, and for the third area, information extraction, feedback to the general framework will be sought.

By combining evaluation of products and projects, the project will gather experience in all three types of evaluation: progress evaluation, diagnostic evaluation and adequacy evaluation.

Exploitation and Future Prospects

Evaluation is a key component of every technology, in that it allows to assess the performance of systems using a given technology and to match system performance against user requirements. In the field of NLP, assessment is one of the most crucial horizontal actions, as positive results in this area can have a most beneficial impact on the entire field.

The TEMAA project will pave the way towards the creation of a set of commonly agreed methods for the evaluation of NLP systems, components and techniques, by

- providing a general framework for evaluation
- using this framework in a concrete way for a few different types of NLP systems.

The benefits of having such a framework are obvious: for industry, the framework may be used to guide development, to measure progress and adequacy, to give diagnostics for systems under development; for decision makers in charge of procurement, it can be used to guide decision; for the research community, it can be used for guiding research, as well as for progress and diagnostic evaluation.

Contact Point

Mrs Bente Maegaard
Center for Sprogteknologi
Njalsgade 80
DK-2300 Kobenhavn S
Denmark

Tel: +45 31 54 22 11

Fax: +45 31 54 61 97

e-mail: bente@cst.ku.dk

Partners

Center for Sprogteknologi (Coordinator) (DK)

ISSCO (CH)

Stichting Taaltechnologie (STT) (NL)

Claris (IRL)

Start Date:	January 1994
Duration:	24 months
Resources:	65,5 person-months
Estimated cost:	604.076 ECU

Interactive Corpus-based Translation Drafting Tool (TRANSLEARN)

Objectives

The aim of the project is to provide a computational methodology and, in more practical terms, a toolbox which will aid the human translator working in a particular subset of general language (a sublanguage) in the following two ways:

- relieve him from the repetitive part of his work, mostly dealing with specialised types of text;
- to enhance productivity and translation quality by assisting him through proposed alternative solutions as well as providing sophisticated ancillary tools.

A prototype application demonstrating the validity of the approach and allowing it to be evaluated in terms of translator productivity will be produced as a result of the project. The project will initially consider four languages: English, French, Greek and Portuguese.

Approach and Methodology

TRANSLEARN is based upon sophisticated pattern matching techniques, involving both linguistic and statistical processing, which are used to identify the longest coherent part of source text which has already been translated and stored in a text database in both source and translated form. In the case of a full match between a piece of source text and a database entry, the corresponding translated text can be output automatically. Statistically ranked alternative translations can also be provided, if they exist. If no full match is detected, a reconstruction and optimal evaluation of all the partial matches is performed which is then, together with a confidence measure, presented to the translator. Fragments of source text for which translations above a certain confidence threshold do not exist will be presented to the translator for him to translate. The translation is then incorporated into the database for future use.

Existing field-proven techniques and utilities will be used for the creation of the database of parallel texts.

TRANSLEARN will collect and investigate a large body of translated texts within a well-defined sublanguage and text type, including the EC CELEX database, select the most coherent and homogeneous set of standard texts, and store these in an appropriately designed text database using existing software text handling and alignment tools. A linguistically and statistically-based pattern-matching mechanism, to be triggered by a source text, will then be developed. The most frequently used fixed locutions and syntactic structures in the sublanguage considered will be stored in a separate database, as will statistical data concerning the text database.

Maximum use of existing products and software techniques will be made, and the sublanguages used for the prototype will be from administrative (EC regulations etc.) and technical (software documentation) texts. The prototype will be limited to fairly simple morphological and syntactic processing, and to known statistical techniques for clustering and taxonomy derivation for fixed locutions.

Exploitation and Future Prospects

TRANSLEARN attempts to combine the statistical and symbolic/knowledge-based approaches to NLP, which are often regarded as mutually incompatible, in a synergistic way, and produce a large database of appropriately organised, indexed parallel texts in two sublanguages in an easily accessible form. The prototype software package produced will be a powerful tool for pattern-matching and other intelligent applications. Tools of this kind are expected to turn into highly marketable products, and TRANSLEARN will be marketed both as a stand-alone utility and as an integral part of a toolbox with wider scope. It is intended to eventually extend the prototype to cover the remaining EC official languages, and to get feedback on its functionality from translation services dealing with the types of text covered by the project.

Progress

By November 1993 TRANSLEARN was working on text corpora in the four languages from the CELEX database, in total 180 MB, i.e. between 5 and 6 million words per language. These corpora are being cleaned for information that cannot be exploited by the project, tagged and lemmatized. The corpora are also being aligned, so that each paragraph and sentence in the Greek, Portuguese and French corpora is linked to the corresponding paragraph and sentence in the English corpus. This work is almost finished and provides the basis for testing and developing the pattern matching techniques and the retrieval of already translated text in the following phases of work.

Contact Point

Stelios Piperidis
ILSP
22 Margari Street
115 25 Athens
Greece

Tel: +30 1 6712 250

Fax: +30 1 6471 262

Partners

Institute for Language and Speech Processing (coordinator) (GR)

University College London (UK)

Instituto de Linguistica Teorica e Computacional (PT)

Sonovision Itep Technologies (F)

Knowledge A.E. (GR)

Start Date:	January 1993
Duration:	30 months
Resources:	95 person-months
Estimated cost:	750.000 ECU

Creation, Reuse, Normalisation and Integration of Terminologies in Natural Language Processing Systems (TRANSTERM)

Objectives

In the domain of technical documentation terminology is probably the most important ingredient. Vast amounts of terminological resources are available for the traditional disciplines of science and technology and new ones are continuously created, enlarged and up-dated for the emerging new disciplines. It is a great challenge for language technology to exploit existing terminologies and to devise tools which support the creation of new data and resources. TRANSTERM addresses the problems of enriching terminologies and integrating them into the application dictionaries of NLP systems. It also deals with automatic and semi-automatic construction of application terminologies from corpora. The main objective is to facilitate the use of terminological data in NLP systems thus tackling the critical issue of real site customisation of this type of software. Two classes of users are foreseen, namely application developers and terminology builders/administrators.

Approach and Methodology

There are three major lines of action:

- The elaboration of a standardised generic representation of terminological data enriched with linguistic information, of and application specific knowledge derived from terminological resources.
- The implementation of a modular portable toolbox allowing a) the assembly and customisation of terminological resources in order to characterise and enrich these resources, check their coherence and merge them with lexical data to create machine-processable lexico-terminological objects and b) semi-automatic terminology extraction from text.
- The validation of the tools, methods and formats developed within the project by means of three real site tests involving corporate data and two smaller-scale experiments covering altogether five languages (French, Italian, English, Greek and Portuguese).

The project is based on methods and tools already existing within the consortium, or under development. Results from related EC sponsored projects (e.g. ESPRIT projects TWB-II, ACQUILEX and MULTILEX) and from the EUREKA projects GRAAL and GENELEX will be used. TRANSTERM is complementary to GRAAL and GENELEX, which deal with the generic grammatical and lexical components of NLP systems.

The TRANSTERM toolbox will also take into account the known document description means (such as SGML) in order to facilitate both the acquisition and reuse of terminological data. Existing international norms in the field of terminology will be taken into account and links will be established with ongoing standardisation efforts in this field (like LISA's TIF) and neighbouring areas (e.g. the Knowledge Interchange Format). The software will be developed on a UNIX platform considering emerging standards such as OSF/Motif.

Exploitation and Future Prospects

The project is very much user driven. The industrial consortium members expect to improve the productivity of their applications, especially in the area of automatic indexing. The software toolbox will allow the construction of application specific disambiguation heuristics and mappings of identified grammatical constructs on to objects conforming to the characteristics of a terminology.

Semi-automatic construction of terminological resources in languages such as Greek and Portuguese will be supported by providing tools usable in these environments.

TRANSTERM is expected to lead to pre-industrial prototypes which lend themselves to rapid exploitation by industrial system developers leading to marketable products. Associated services will become more cost-effective. The results of work on standardisation will be made available to the scientific and industrial communities.

The close cooperation of TRANSTERM with the related Eureka projects GRAAL and GENELEX will have a synergetic effect on Community sponsored efforts in Natural Language Processing.

Contact Point

Mr Antoine Ogonowski
GSI-ERLI
1, Place des Marseillais
F- 94227 Charenton Cedex
France

Tel:	+33 1 48 93 81 21 (secretary)
Fax:	+33 1 43 75 79 79
e-mail:	Antoine.Ogonowski@erli.gsi.fr

Partners

GSI-ERLI (Coordinator) (F)

Electricité de France (F)

Aèrospatiale (F)

Institute for Language and Speech Processing (GR)

Centro Ricerche FIAT (I)

University of Surrey (UK)

Associated Partners

ISSCO (CH)

LINGSOFT (FI)

ILTEC (PT)

ITC/IRST (IT)

Start Date:	December 1993
Duration:	24 months
Resources:	172 person-months
Estimated cost:	1.949.025 ECU

Test Suites for NLP Applications (TSNLP)

Objectives

As the market for NLP products and services is growing, clear patterns of types of systems start to emerge. In the future, prospective buyers and end-users of NLP products will be confronted more and more with the problem of choosing the product that best meets their specific requirements. Suppliers of NLP products and services may want to know how their systems and tools compare to those of their competitors. Developers and researchers are likely to be interested in figuring out whether their system performs according to their specifications. At present, companies, institutions and corporate users interested in any of the above-mentioned evaluation types spend a considerable amount of time and effort in building data and tools for their own test purposes. This project aims to alleviate the situation with respect to data and tools, by defining guidelines and a methodology for the construction of diagnostic data ("test suites") and designing and implementing related tools.

The guidelines will be validated by constructing application-specific diagnostic data for French, German and English and testing them on a number of applications ranging from parsers through grammar checkers to controlled language checkers. In addition to devising guidelines, the project is to investigate techniques and design and implement tools that will facilitate the construction, use and manipulation of test suites, such as a database for storing and manipulating test data, and tools for the (semi-) automatic generation of test suites.

Approach and Methodology

The project starts out with a survey of existing tests suites. This survey will help identify the types of NLP applications for which a given test suite has been used, and the type of evaluation where it has been applied. This preparatory work will serve as a starting point for the definition of guidelines for the construction of test suites, which will be undertaken in the second, central workpackage. Some of the issues which the project will investigate may be application-independent (e.g. size), others may be application-dependent (e.g. the necessity of avoiding examples which involve translational problems when constructing a test suite for monolingual applications). The project will also investigate ways of assigning weightings to test sentences and an efficient annotation scheme. The annotation scheme adopted will be developed with the aim of storing the test suite fragments in a database.

The soundness of the proposed guidelines and methods will be demonstrated by constructing test data for French, German and English, and validating the resulting test suites against a number of applications and/or components.

The project intends also to investigate techniques and design and implement tools that will facilitate the construction/generation, use and manipulation of test suites.

Firstly, the project will investigate techniques for automatically generating test suites, e.g. by means of special, simple test suite grammars. Test suites are normally hand-constructed. However, this process is difficult (requiring considerable linguistic sophistication and skill), laborious, tedious, and above all error prone. All this suggests that the process is a good candidate for automation, or more precisely, for an interactive process that involves a substantial amount of automation. Another advantage of automation is that this allows "dynamic" test suite construction, where test data can be replaced by new data which test the same phenomena. In this way, it may be possible to overcome one of the problems that sometimes crop up in system evaluation, namely that developers can tune their application so that it deals with static test data. Finally, automatic, dynamic test suite generation should open the possibility of using very large lexicons, perhaps with some "randomisation", thus making it possible to hold and transmit extremely large 'virtual' suites, in the order of many millions of sentences, providing standard benchmarks for system testing.

Secondly, the project will investigate whether and to what extent it is possible to derive test suites (semi-) automatically from corpora.

Finally, the project intends to design and develop a relational database for storing and manipulating test suite data. The annotated test suite fragments built during the project will be stored in that database.

Exploitation and Future Prospects

The main result of this project will be the guidelines and the methodology for test suite construction, that can be used in different NLP application fields and systems. It is expected that a set of guidelines will facilitate the interpretation of test suites and enhance their portability. This will be of direct benefit to all those companies and institutions that nowadays spend a considerable amount of time and effort in building test suites for their own purposes.

The results are also likely to be useful for several areas of linguistic research, since they provide a catalogue of linguistic data of potential value to theoretical and empirical work in linguistics.

All project results will become publicly available. The tools will have a high degree of portability, allowing for easy integration into a common framework (e.g. ALEP). The availability of projects results will be widely publicised at conferences, evaluation workshops and in magazines in the field of NLP, in order to create optimal conditions for exploitation by a wide number of users.

Contact Point

Mrs Siety Meijer
Mr Doug Arnold
CL/MT Group
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
United Kingdom

Tel: +44 206 872084 / 872086

Fax: +44 206 872085

e-mail: meyes@uk.ac.essex

Partners

University of Essex (Coordinator) (UK)

ISSCO (CH)

Aerospatiale (F)

DFKI (D)

Start Date:	December 1993
Duration:	20 months
Resources:	72 person-months
Estimated cost:	575.700 ECU

For more information about the LRE sub-programme, contact:
Robert Cencioni (programme manager)
EC, DG XIII E-4, Bâtiment Jean Monnet, Plateau du Kirchberg, L-2920 LUXEMBOURG
Telephone: +352 4301 32886 Fax: +352 4301 34999