

DELIVERABLE IDENTIFICATION

Identification number	LRE-63314-D3.2.2
Type	Technical Report
Title	Relations of a European Center for Spoken Language Resources (ECSLR) with on-going projects
Status	Final
Deliverable	D3.2.2
Work Package	WP 3
Task	Task 3.2
Period covered	Febr 1995 - March 1995
Date	31 July 1995
Version	Final
Number of pages	15
Author(s)	J. Mariani (LIMSI-CNRS)
Work package (WP) / Task (T) responsible	WP3: H. Höge, J.-M. Dolmazon / T3.2: J. Mariani
Project contact point	Harald Höge, Siemens AG, ZFE T SN 5, D-81730 München Phone: +49 89 636 53374, Fax: +49 89 636 49802 E-mail: hoege@habicht.zfe.siemens.de
CEC project officer	José Soler
Status	Public
Actual distribution	Consortium and CEC
Supplementary notes	
Key words	
Abstract	
Status of abstract	

Received on	
Recipients catalogue number	

DOCUMENT EVOLUTION

Version	Date	Status	Notes

Mlap SPEECHDAT project
Deliverable 3.2.2.

Relations of a European Center for Spoken Language Resources (ECSLR)
with on-going projects.

J. Mariani
LIMSI-CNRS

I. Introduction

The decision was to create a common Language Resources (LRs) distribution entity for the different types of LRs, including the Spoken Language ones. The European Language Resources Association, ELRA, has been created as a joint effort of the LRE project RELATOR, and of the 3 Mlap projects Speechdat, Parole and Pointer, with the support of the Commission of the European Community. Its goal is to gather linguistic resources, especially for the languages spoken in Europe, and to make them easily available, especially to the European companies and reeseach laboratories in order to develop and test Language Engineering systems, and conduct research aiming at such systems.

Its relationship with on-going projects must be established. First with the on-going CEC projects, and especially the LRE Relator project, and the 3 Mlap projects. But also with the ESPRIT BRA European Language and Speech network ELSNET. With the CEC projects in different programs (mostly Esprit and LRE) which produced, produces or will produce Linguistic Resources of interest, that could be distributed through the ELRA. An extension to this relationship to Eastern European languages will be made possible through existing Copernicus projects covering spoken (BABEL) or written (TELRI) LRs for the corresponding languages in Eastern Europe. The same apply with other cooperative research projects, such as the Eureka program. Another link should be established in the field of standards and normalization of LRs with the LRE Eagles project. And in the field of assessment with any initiative taken by the CEC to cover this area within the 4th Framework Program. Other attention should be devoted to the relationship with projects at the National level, especially in Europe. At the international level, the relationship with the LDC, both in terms of gaining from their 5-year experience in the field of LRs distribution, and on the share of efforts for distributing data is an important issue. Finally, ELRA must be an important actor in the international bodies addressing the field of LRs, such as Cocosda for spoken language resources (SLRs), and Liric for written ones (WLRs).

II. The LRE Relator project

The LRE Relator project has been initialized by the "Resource Reusability" Task Group of the ELSNET Esprit BRA Network of excellence in Language and Speech. It has 8 participants (University of Pisa as prime partner, University of Copenhagen (Denmark), Institut de recherches comparatives sur les Institutions et le Droit of CNRS (France), LIMSI-CNRS (Orsay, France), INESC (Lisbon, Portugal), DFKI (Sarrebrucken, Germany), University of Stuttgart (Germany) and Center for Cognitive Sciences (Edinburgh University (UK)). The total effort is about 55 Man-Months. The project is organized as 5 Working Groups:

WG1. Evaluation of the needs of LRs in Europe

This WG addresses the various types of LRs: text corpus, speech corpus, lexical and terminological data, grammars and tools. For each type of LR, it studies the various aspects of resources: taxonomy, status, availability, necessary effort to produce LRs of a given type.

WG2. Definition of possible organizational models

This WG studied the possible organizational models for the setting of a LR infrastructure in Europe, taking into account the various actors (and especially the industrial ones), the European National efforts and the legal aspects. In order to conduct this study, the WG settled an Industrial Steering Committee, chaired by V. Parajon-Collada, Deputy General Director of DG XIII at the CEC. This WG prepared the launching of the European Language Resources Association (ELRA).

WG3. Liaison with non-EU entities

This WG studied the existing LRs outside the EU, and the possible cooperation modalities with the corresponding countries (Eastern and Central European ones and others). Two projects have been launched within the Copernicus program (BABEL project for spoken language and TELRI project for written language). Contacts are also taken with the US (LDC), with international Committees (Cocosda for SL and Liric for WL), and with comparable programs in various countries (Japan, Australia, China, Korea etc). The international liaison for the Spoken Language part is supported within the Eurococosda LRE project.

WG4. Experimental distribution of LRs through CDRoms.

The use of CDRoms as the distribution media has been studied. Various aspects have been studied: acquisition of existing data (data identification and selection, distribution agreement), production of CDRoms (verification of data and standardisation), CDRom pressing and distribution. This will give an evaluation of the costs of such distribution, which is especially interesting for SLRs, as they need a lot of space.

WG5. Experimental distribution through the network.

Another distribution media is also studied. The operations are similar (LRs acquisition, distribution agreement, verification of data and corresponding documentation, evaluation of technical feasibility and cost) The AFS software has been experimented (3 servers have been installed in Edinburgh, Sarrebrucken and Pisa). This media is mostly interesting for WLRs, as they don't occupy much space.

ELRA has been conceived by Relator, and will gradually cover the activities that Relator covers presently in an experimental way. The legal aspects of LRs distribution have been studied within Relator, but will need a much larger attention within ELRA.

III. The 3 MLAP projects for Spoken, Written and Terminological Linguistic Resources production

The distribution entity should be closely linked to the production infrastructure. In the present MLAP program, 3 projects are discussed that deal with the production of linguistic resources: Speechdat, which aims at the production of spoken language resources, Parole which aims at the production of textual resources (lexica and corpora), and Pointer which addresses the terminological resources. As far as spoken resources are concerned, Speechdat considers 4 different areas of actions: i) the production of short-term databases (databases that are of direct, short term interest for the European telephone industry). In this framework a case study database will be produced. ii) the production of long-term databases (those which are necessary to develop future advanced speech technology and to support speech research) iii) Working standards and tools, in order to ensure the easy sharing and reusability of resources, and the assessment of the quality of the databases. Such tools may comprise DBs headers, quality insurance checkers, compression, scoring and labelling softwares, signal display... A specific subtask also addresses the definition of databases aiming at the assessment of automatic speech processing systems. iv) the

definition of a distribution infrastructure and the relationship with other projects, both from the CEU, from Eastern European countries and from the international scene. This item is directly related to the object of Relator and the efforts have been coordinated, as it appeared that there was a need for both Spoken and Written Language Resources distribution, resulting in the launching of ELRA.

Speechdat will focus accordingly to the production, and validation, of SLRs, and will make available through ELRA the SLRs that will be produced within Speechdat as a Case Study.

Two other corresponding projects exist for WLRs and Terminology: Parole and Pointer. They have the same relationship with the ELRA, and will also focus on LR's production. The Parole project has established a Belgian non-profit Association to coordinate the WLRs production of major European Institutes.

IV. The European Language and Speech Network (Elsnet)

The European Language and Speech Network (Elsnet) has been launched within the ESPRIT BRA program in April 1991. It includes about 50 public research laboratories, and about the same number of industrial affiliate members. The network focus its activities on the integration of language and speech. The "Resource Reusability" Task Group (RRTG) is headed by A. Zampolli. It has already distributed several SLRs corpus through CDRoms (the Dutch "Groningen" corpus, the French "Bref-Polyglot" corpus). Elsnet financed the formatting of the corpus by its producer and the CDRom pressing. The sales are shared between the producer and Elsnet, in a non-linear way (initially 80% to Elsnet until the break-even point has been reached. Then, it goes up to 80% for the producer). In the case of WLRs, the use of electronic network for distribution has been studied. Elsnet also sponsored the production and distribution of the ECI ("European Corpus Initiative") CDRom, which includes several text corpus in various European languages and is also distributed by the LDC. The ECI has 46 corpus in 27 different languages, for a total of 92 millions words, from various sources (newspapers, broadcasting transcriptions, lexica, technical texts, conference proceedings etc). Elsnet recently received a Copernicus contract to extend to Eastern countries, and an INTAS contract to extend to FSU countries, and may also include LR's production and distribution within these contracts.

Elsnet played a major role in the launching of the ELRA, through the Relator project. The role of the RRTG will change with the existence of the ELRA. It will probably focus on experimental activities, such as annotation of LR's, or experimental production of LR's for specific needs. The continuing role of Elsnet as a distribution entity will depend on how fast the ELRA will function practically and, in the long term, on the response of the ELRA to the needs of the European Speech and NL research community.

V. The past or on-going CEC projects where LR's have been or are produced

Many projects have produced LR's in the past Esprit or LRE programs (see Annex 1 & 2). This data may be available, at a minimum cost. The availability, quality and cost of such data has to be checked. Some projects are especially in a position to offer such data: the Esprit SAM (Eurom 0 and 1 multilingual speech data), Polyglot, Sundial, Sunstar and Accor for SLRs, the Esprit Multilex and Esprit BRA Acquilex for Lexica. The LRE projects SQALE, TEMAA and TSNLP for data related to the assessment of LE systems, Onomastica for proper nouns multilingual (11 languages) lexica, Eurococosda for speech in English pronounced with various accents and through different medias (2 microphones, a laryngophone and the telephone) (TED), Translearn for aligned texts and Transterm for Terminology. The LRE Multext project, and its Eastern European extension, will also be a major source of multilingual annotated (TEI/Eagles format) WLRs. Other possible sources are the Delis, LS-Gram, RGR and SIFT projects. In the future, it is desirable that the data produced within the new FP4 Telematics/LE program will be distributed by the ELRA.

VI. The Copernicus projects and other projects

The Copernicus program aims at the cooperation with Eastern and Central European countries. The BABEL project has been started, in close conjunction with the LRE Eurocosda project. It deals specifically with the production of SLRs in various Eastern European languages (Bulgarian, Estonian, Polish, Hungarian and Romanian). The project is coordinated by the University of Reading (UK). The targeted data will be in agreement with the Esprit SAM recommendations, and a CDROM will be edited for each of the 5 languages and will be distributed through the ELRA. A comparable project (TELRI) has been also started for WLRs.

Other data can possibly be obtained through Eureka projects, and especially through Genelex (Lexica), Graal (Grammars) and EuroLang (multilingual (10 languages) data for Computer Aided Translation).

VII. The LRE Eagles project

The LR distribution infrastructure should have close links with any action dealing with standards that may be common, or specific, to speech and/or text. This is presently covered by the LRE Eagles (Expert Advisory Group for Language Engineering Standards) project. Eagles is organized as 5 Working Groups, each with a host, a chair and an editor to produce the documents:

WG	Host	Chair	Editor
- Text Corpus	Inst. Cervantes Madrid (Spain)	A. Zampolli	J. Llisterra
- Lexica	GSI-ERLI Paris (France)	A. Zaenen	U. Heid
- Evaluation	Center for Sprogteknologi Copenhagen (Denk)	M. King	C. Mazzi
- Formalisms for computational linguistics	DFKI Sarbrücken (Germany)	H. Uszkoreit	S. Spackman
- Resources for Spoken Language	Vocalis Ltd Cambridge (UK)	R. Moore	D. Gibbon

The standards should cover very large gata bases (SLR, WLR and lexica). It should address the way to manipulate data and to validate resources, tools and products. About 100 sites participate in Eagles. The tasks of each WG is the following:

WG1. Text Corpus

Establish a typology of corpus, propose Corpus Encoding Standards, as an extension of TEI (Text Encoding Initiative), define linguistic annotations (morphosyntactic and syntactic).

WG2. Computational Lexica

Three tasks are planned: Validation and consolidation of the work already conducted on morphosyntax, syntactic encoding and preparation of other aspects of lexica standardisation.

WG3. Evaluation

The task here is to provide guidelines in this area: how to define a user "profile", how to test grammar checker, how to test Machine Translation systems, with the CEC translation services as a case study.

WG4. Formalisms

A basic study is conducted here on the fundamentals of computational linguistics and on how they are related to standards.

WG5. Spoken Language

This task is subdivided in 4 sub-tasks: voice input, corpus (physical characterisation and description, "format & tools" and "design methodology").

A huge report is to be produced in Fall 1995.

Obviously, a close relationship should be established between ELRA and EAGLES, or what will follow Eagles, in order to distribute data which is in agreement with the recommendations made up by the experts in the various fields. ELRA would thus put in practice, or ask the validation units to put in practice, the results of the Eagles WG studies. It should also take care of the distribution of tools (for speech or text compression, annotation, for speech visualization...). Those tools can be obtained from laboratories, but may also be designed at the ELRA.

VIII. Relationship with Assessment

The availability of LRs for LE systems assessment is of great importance. The data can be made available to all participants in an evaluation exercise, in the same way and in agreement with the evaluation protocol (making first the training data available, then the data for the dry-run test and finally for the test itself). This data must be produced cautiously: for example, the producer shouldn't be also the tester of its own material. A centralized entity can take care that the necessary precautions have been undertaken. Also, the data is available afterwards for any laboratory or company wishing to check in an informal way the quality of its system or its product to the results which have been reported after the evaluation test on this data. It will be important to establish the relationship between ELRA and any entity that may be created (within the CEC or outside) for addressing spoken and written language processing systems assessment.

IX. ELRA and European National efforts

At the European level, several countries have produced, or produce LRs that could be distributed by the ELRA. In UK, several corpus have been produced (WJSCam, Maptask, British National Corpus, Cambridge Language Survey, Longman/Lancaster corpus, London/Lund corpus, Bramshill corpus...). In France, several corpus have also been produced: BREF (very large newspaper texts read aloud), BDSons (basic French Speech), BDBruit (noisy speech), BDLex (lexica). Now a large program on system evaluation, including the production and distribution of French language resources has been started by the Aupelf-Uref, for Document Retrieval, Text Alignment, Terminology DB Extraction, Voice Dictation, Vocal Dialog and Speech Synthesis. Aupelf and CNRS are also starting a LRs server network. And CNRS is starting an evaluation exercise on morphosyntactic taggers, including training and test material (GRACE project). In Germany, the large Verbmobil project produces a very large corpus of vocal dialogs, and other data is already available (Phondat). Spex in The Netherlands produces large corpus in Dutch. The Dante corpus has been produced in Italy as a joint effort between research and industry, sponsored by the Post Office. In Spain, the Albayzin corpus has been produced. Other WLRs or SLRs corpus can be found in Belgium (Beltex...), Greece (ILSP...), Portugal (CRPC...), Sweden (Waxholm...), Denmark (CPK...) etc. Large efforts on the production of

LRs have been devoted in National Institutes: INALF in France (Trésor de la Langue Française, Frantext, DELA lexic...), IDS in Germany, INL in The Netherlands etc.

X. ELRA and the LDC

LDC already exists since 1991. It has European members and distributes European LRs (see Annex 9). The links between ELRA and LDC must be established in order to have a good cooperation between the two entities and avoid competition and redundancy. LDC seems to mainly target the US market, not the international one. They would be pleased to be able to distribute data in European languages to their US members, but would probably be glad to allow the US data to be distributed in Europe through the ELRA. This cooperation agreement has to be carefully designed, taking into account the fact that the US industry needs data in various languages to attack the corresponding markets, but that LDC presently distributes their data in American English to anybody wishing to become a member, without any apparent protectionism.

XI. International liaison

Finally, ELRA should also have links with other actors on the international scene. The International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment (Cocosda) exists since 1991, and has participants from all continents (see Annex 5). The interface of European laboratories with Cocosda is partly supported by the CEC (LRE Eurococosda project). A workshop is to take place after Eurospeech'95 in Madrid (September 1995), where the ELRA will be introduced on the international scene. The equivalent of Cocosda for written language (Liric) is presently worked out. Important LRs production projects can also be found in Japan (EDR, ATR, JEIDA, Voice Across Japan (TI)...), Korea (KCCSLP), Australia (ANDOSL...), China (CAS) etc...

XII. References

CEC reports:

- A. Danzin, "Vers une infrastructure linguistique européenne", rapport pour la DXIII de la CCE, Mars 1992

French reports:

- A. Danzin, "Le Français soumis au choc des Technologies de l'Information. Propositions pour une politique offensive", Rapport pour les Ministères de l'industrie, des Postes et Télécommunications et du Commerce Extérieur, de la Culture et de la Francophonie, et de l'Enseignement Supérieur et de la Recherche., Décembre 1994

- "L'avenir de la Langue Française, synthèse des débats", Commission de la République Française pour l'Education, la Science et la Culture (UNESCO), Fondation Singer-Polignac, Janvier 1993

- A. Abbou, "Le Multilinguisme en Europe Communautaire", Rapport pour le MESR-DIST, Décembre 1994

- R. Carré, R. Descout, J. Mariani, M. Eskénazi, M. Rossi, "The French Language Database : Defining, Planning and Recording a Very Large Database.", IEEE ICASSP Conference, San Diego (USA), March 19-21 1984

- G. Pérennou, "BDLEX: a Data and Cognition Base of Spoken French", ICASSP'86, Tokyo, 7-11 Avril 1986

- I. Ferrane, M. de Calmes, D. Cotto, J.M Pecatte, G. Pérennou, "Analyse lexicale de la base de données BREF", 19èmes Journées d'Etudes sur la Parole, Bruxelles, 19-22 Mai 1992

- G. Pérennou, D. Cotto, M. de Calmes, I. Ferrane, J.M Pecatte, "Le projet BDLEX de base de données lexicales du Français écrit et parlé", Séminaire Lexique du PRC Communication Home-Machine, Toulouse, 21-22 Janvier 1992

- L.F. Lamel, J.L. Gauvain, M. Eskénazi, "BREF: a Large Vocabulary Spoken Corpus for French", Eurospeech'91, Gênes, 24-26 Septembre 1991

- B. Courtois, M. Silberstein, "Dictionnaires électroniques du Français", Revue Langue Française N°87, Larousse, Sept. 1990

- E. Laporte, "Le Dictionnaire phonémique DELAP", Revue Langue Française N°87, Larousse, Sept. 1990

- M. Silberztein, "Le dictionnaire électronique des mots composés", Revue Langue Française N°87, Larousse, Sept. 1990
- J.P. Tubach, L.J. Boë, "De A à Zut-Petit Dictionnaire Français", Rapport interne ICP, 1985

NSF/CCE reports

- R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, A. Zampolli, V. Zue, 'A Survey of the State-of-the-Art in Human Language Technology', Rapport NSF à paraître Juin 1995

Cocosda proceedings

- K. Jones, J. Mariani Eds, "Proceedings of the 1992 Workshop of the International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment", Banff Springs Hotel, Octobre 1992
- K. Jones, A. Fourcin Eds, "Proceedings of the 1993 Workshop of the International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment", Haus am Köllnischen Park, Berlin, Septembre 1993

Eagles reports:

- J. Llisterri, "Introduction to the reports of the Text Corpus Working Group"
- J. Sinclair, "Corpus Typology"
- N. Ide, J. Véronis, "Corpus Encoding"
- G. Leech, A. Wilson, "Morphosyntactic Annotation"
- J. Llisterri, "Spoken Texts"
- U. Heid, "Synthesis of Computational Lexicons Working Group"
- B. Menon, N. Modiano, "Task Group on lexicon architecture"
- M. Monachini, N. Calzolari, "Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora"
- M. King, "Evaluation of NLP Systems"
- R. Backofen, H.U. Krieger, S.P. Spackman, H. Uszkoreir, "Report of the Eagles workshop on Implemented Grammar Formalisms"
- L. Boves, D. Van Bergem, Els den Os, "Spoken Language Corpora"

Relator reports:

- A. Bech, U. Heid, E. Hinkelman, L. Lamel, I. Trancoso, J. Villadsen, J. Zeiliger "Report on European NLP Resources"
- L. Lamel, I. Trancoso, J.M Dolmazon, "Survey of Spoken Language Resources"
- J. Villadsen, A. Bech, "Survey of Grammars and Related Tools"

ESPRIT projects reports

- J. Mariani, "An overview of ESPRIT programmes on Speech and Natural Language Processing", in A. Varghese, J.P. Lefebvre eds, Esprit Speech Project Book, Springer Verlag, 1994
- A.J. Fourcin, G. Harland, W. Barry, V. Hazan Eds, "Speech Input and Output Assessment. Multilingual Methods and Standards", Chichester: Ellis Horwood Ltd, 1989
- "European Summer School on Language and Speech Communication: Corpus-Based Methods", Research Institute for Language and Speech (OTS), Université d'Utrecht, 11-22 Juillet 1994

XIII. Annexes

Annex 1: Esprit Projects related to LRs and evaluation

Annex 2: Linguistic Research and Engineering (LRE) projects related to LRs and evaluation

Annex 3: Mlap Projects related to LRs and evaluation

Annex 4: Eureka Projects including LRs

Annex 5: Cocosda Members

Annex 6: Relator Industrial Steering Committee members

Annex 7 : Plans for selling Spoken Language Resources

Annex 8: Presentation of the LDC; subscription modalities and sale of data (1995)

Annex 9: Data available at LDC (1995)

Annex 1: Esprit Projects related to LRs and evaluation

- **SAM 1 (MULTILINGUA): Multilingual Speech Input-Output Assessment Methodology and Standardisation**
Univ. College of London (UK), CSELT (Italy), JYSK Telefon (Denmark), Amsterdam Univ. (Netherl.), CNET (France), then Univ. College of London (UK), NPL, Smiths Industries, RSRE, Logica (UK), CSELT (Italy), JYSK Telefon (Denmark), TNO-Amsterdam Univ., Dr Neher Lab. (Netherl.), GRECO (CNET, CRIN, CERFIA, ENST, LIMSI, IPA, ICP) (France), duration: 12 months

- **MULTILEX: The Multi-Purpose standard lexicon**
Triumph-Adler (Germany), CAP-SESA, ASSTRIL, IMAG (France), Univ. Pisa, Lexicon (Italy), L-Cube (Greece), Philips, Vrije Univ. Amsterdam (Netherl.), Siemens-Nixdorf, Univ. Munster, RUB (Germany), Univ. of Surrey, Univ. of Manchester (UK), Systems Spain (Spain), Duration: 36 months

- **EMIR: European Multilingual Information Retrieval**
CEA-CEN, SYSTEX (France), Un. de Liège (Belgium), Transmodul (Germany), Duration: 36 months

- **ACQUILEX: Acquisition of Lexical Knowledge for Natural Language Processing Systems**
CNR/Un. di PISA (Italy), Un. Amsterdam (Netherl.), Un. Cambridge, Cambridge Univ. Press (UK), Un. College Dublin (Ireland), Un. Politecnica de Cataluna (Spain), Duration: 30 months

- **449: Investigation into the effective use of speech at the human-machine interface**
British Maritime Technology, Voice Systems International, ICL (UK), Fincantieri (Italy), Duration: 13 months

- **SUNDIAL: Speech Understanding and Dialogue**
Logica, Univ. Surrey (UK), AEG, Erlangen Univ., Siemens (Germany), CNET, IRISA, CAP-SESA (France), SARIN, CSELT, Politecnico di Torino (Italy), Duration: 60 months, 154 MY

- **SUNSTAR: Integration and Design of Speech Understanding Interface**
JYSK Telefon (Denmark), Stuttgart University IAT, Fraunhofer Gesellschaft IAO (Germany), Telefonica (Spain), Alcatel-Face (Italy), INESC (Portugal), Duration: 60 months, 166 MY

- **POLYGLOT: Multilanguage Speech-to-Text and Text-to-Speech system**
Olivetti, Software Sistemi (Italy), LIMSI, BULL (France), Ruhr Univ., Triumph-Adler, Philips-Kommunikations, Siemens (Germany), Madrid Univ., UNED (Spain), Patras Univ. (Greece), CSTR (UK), Nijmegen Univ., Philips-IPO (Netherl.), Duration: 36 months, 140 MY

- **SAM 2: Speech Assessment Methodology**
Univ. College of London (UK), NPL, Smiths Industries, RSRE, Logica (UK), CSELT, CNR, Fondazione Hugo Bordoni (Italy), JYSK Telefon (Denmark), TNO-Amsterdam Univ., Dr Neher Lab. (Netherl.), GRECO (CNET, CRIN, CERFIA, ENST, LIMSI, IPA, ICP) (France), AEG, RUB, Bielefeld Univ. (Germany), Televerket, KTH (Sweden), ELAB (Norvège), Duration: 36 months

- **ACCOR: Articulatory-Acoustic Correlation in Coarticulatory Processes: a cross-Language Investigation**
IPA-CNRS (France), CNR-Padova (Italy), Siemens, Maximilians Univ. (Germany), Un. Valencia, Un. Politecnica Barcelona (Spain), Un. Reading (UK), Un. Dublin (Ireland), Univ. Stockholm (Sweden), Duration: 30 months

- **ELSNET: Network of Excellence on Language and Speech**
CCS (Edinburgh Univ., UK), Un. Dublin (Ireland), OTS-Utrecht, Amsterdam Univ. (Netherl.), INESC (Portugal), KTH (Stockholm, Sweden), Pisa Univ. (Italy), Roskilde Univ. (Denmark), Stuttgart Univ. (Germany), LIMSI-CNRS (France)

- **SAM A: Speech Technology Assessment for Multilingual Applications**

Logica, NPL, DRA, (UK), CSELT, Fondazione Hugo Bordoni (Italy), JYSK Telefon (Denmark), TNO, (Netherl.), Vecsys, ENST, ICP (France), RUB, Bielefeld Univ. (Germany), Televerket (Sweden), INESC (Portugal), Patras University (Greece), Polytechnic University of Catalunya (Spain), Duration: 36 months

- **WERNICKE: A Neural Network Based Speaker Independent, Large Vocabulary, Continuous Speech Recognition System**

Lernout & Hauspie (Belgium), INESC (Portugal), Univ. Cambridge (UK)

- **ELSNET: Network of Excellence on Language and Speech**

CCS (Edinburgh Univ., UK) puis OTS-Utrecht (Netherl.) depuis Janvier 1995, Un. Dublin (Ireland), OTS-Utrecht, Amsterdam Univ. (Netherl.), INESC (Portugal), KTH (Stockholm, Sweden), Pisa Univ. (Italy), Roskilde Univ. (Denmark), Stuttgart Univ. (Germany), LIMSI-CNRS (France)

- **ACCOR II: Articulatory-Acoustic Correlation in Coarticulatory Processes: a cross-Language Investigation (Working Group)**

Un. Reading (UK), IPA-CNRS (France), CNR-Padova (Italy), Siemens, Maximilians Univ. (Germany), Un. Valencia, Un. Politecnica Barcelona (Spain), Un. Dublin (Ireland), Univ. Stockholm (Sweden), Duration: 30 months

- **ACQUILEX II: Acquisition of Lexical Knowledge**

Un. Cambridge, Cambridge University Press (UK), CNR/Un. di PISA (Italy), Un. Amsterdam, Van Dale Lexicografie (Netherl.), Un. College Dublin (Ireland), Un. Politecnica de Cataluna, Bibliograf (Spain)

Annex 2: Linguistic Research and Engineering (LRE) projects related to LRs and evaluation

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

Univ. Stuttgart (Germany), SITE (France), Van Dale Lexicografie b.v., Vrije Univ. Amsterdam (Netherl.), Istituto di Linguistica Computazionale (Italy), Center for Sprogteknologie Univ. Copenhagen (Denmark), Lingsoft (Finland, Linguacubun -UK)

- **EAGLES: Expert Advisory Group on Language Engineering Standards**

Istituto di Linguistica Computazionale (Italy), GSI-Erli (France), DFKI (Germany), Center for Sprogteknologie Un. Copenhagen (Sweden), Vocalis Ltd (UK), Instituto Cervantes (Spain)

- **Eurococosda: European interface to Cocosda**

University College of London (UK), LIMSI-CNRS (France), Institute of Phonetics, Amsterdam University (Netherl.), CSELT (Italy), Institute of Phonetics, Munich University (Germany)

- **LS-GRAM: Large Scale Grammars for EC languages**

IAI Sarrebourg, IMS-CL (Germany), Univ. Essex (UK), Fundaciun Bosh Gimpera (Spain)

Multext: Multilingual Text Tools and Corpora

LPL, Eurolangue SITE, Centre de Recherche de Rank Xerox (France), SNI, Munster Univ., CAP Debis (Germany), DEC-NL, Univ. Utrecht (Netherl.), Univ. Pisa (Italy), HCRC/LTG Univ. Edinburgh (UK), ISSCO (Switzerl.), Univ. Autonoma de Barcelona, Central Univ. of Barcelona (Spain)

Onomastica: Multi-language Pronunciation Dictionary of proper Names and Place Names

CSTR Un. d'Edimbourg, British Telecom Laboratories (UK), STC Aalborg, JYDSK Telefon (Sweden), ENST France Telecom (France), Techn. Univ. of Berlin, Deutsche Bundespost Telekom Darmstadt (Germany), Patras University, Intrasoft (Greece), Istituto di Linguistica Computazionale, CSELT (Italy), Catholic Univ. of Nijmegen, PTT Research (Netherl.), INESC, Telefones de Lisboa e Porto (Portugal), Polytechnic Univ. of Madrid, Telefonica (Spain), SINTEF DELAB, Norwegian Telecom Research (Norvège), KTH, Telia (Infovox) (Sweden).

- **Relator: European Network of Repositories for Linguistic Resources**

Istituto di Linguistica Computazionale (Italy), LIMSI-CNRS, ICP (France), Univ. Edinburgh (UK), DFKI, Univ. Stuttgart (Germany), CST Univ. Copenhagen (Denmark), INESC (Portugal)

- RGR: The Reusability of Grammatical Resources

OTS, Utrecht, ITK (Netherl.), HCRC Univ. Edinburgh (UK), Sarrebruck Univ. (Germany)

- SIFT: Selecting Informtion from Text

Univ. Limerick (Ireland), Univ. Amsterdam (Netherl.), Univ. Heidelberg (Germany)

- SQALE: Speech Recognizer Quality Assesment for Linguistic Engineering

TNO-IZF (Netherl.), LIMSI-CNRS (France), Philips-Aachen (Germany), Cambridge Univ. Engineering Department (UK)

- TEMAA: A Testbed Study of Evaluation Methologies: Authoring Aids

CST Univ. Copenhagen (Denmark), ISSCO (Switzerl.), OTS Utrecht (Netherl.), Claris (Ireland)

- TRANSLEARN: Interactive Corpus-Based Translation Drafting Tool

ILSP, Knowledge A.E.(Greece), UCL (UK), Inst. de Linguistica Teorica e Computacional (Portugal), SITE (France)

- TRANSTERM: Creation, Reuse, Normalisation and Integration of Terminologies in Natural Language Processing Systems

GSI-ERLI, EDF, Aérospatiale (France), LSP (Greece), FIAT (Italy), Univ. Surrey (UK), ISSCO (Switzerl.), Lingsoft (Finland), ILTEC (Portugal), ITC IRST (Italy)

- TSNLP: Test Suites for NLP Applications

Univ. Essex (UK), ISSCO (Switzerl.), Aérospatiale (France), DFKI (Germany)

Annex 3: Mlap Projects related to LRs and evaluation

- Pointer

BJL Consult (Belgium), CL servicios Linguisticos, Cindoc, IULA, Termcat, UZEI (Spain), Deutsches Inst. für Terminologie (Germany), INT, CTN, GOTTA, Union Latine (France), Un. Surrey (UK), Infoterm (Austria), Assiterm, EAB (Italy), CRB (Switzerl.), DTG (Denmark), ELOT (Greece), ILTEC (Portugal), TNC (Sweden), Topterm (Pay-Bas)

- Speeechdat

Siemens AG (Germany), Alcatel-Face (Italy), GEC Marconi (UK), Jydske Telefon (Denmark), Philips Research (Germany), Cselit (Italy), Portugal Telecom (Portugal), Vocalis (UK), UCL (UK), Limsi-CNRS (France), Un. Amsterdam (Netherl.), UAB (Spain), Un. Munich (Germany), Un. Aalborg (Denmark), ICP (France), DRA (UK), IDIAP (Switzerl.)

- Parole

Sietec, IDS (Germany), Ist. Linguistica Computazionale Pisa, Societa Generale di Informatica, SOGEI (Italy), Un. Birmingham (UK), Det Danske Sprogog litteraturselskab (Denmark), IDL (Netherl.), Real Academia espanola (Spain), INaLF, Alcatel-CIT, LPL (France), Inst. d'estudis catalans, Grupo CL servicios linguisticos (Spain), Sprakdata, Un. Göteborg (Sweden), St Patrick's College (Ireland)

Annex 4: Eureka Projects including LRs

- Eurolang

SITE, CNET, LADL, GETA, SLIGOS, MATRA Hachette (France), Sietec (Germany), Un. Pisa (Italy), Un. Barcelona (Spain)

- GENELEX

Groupe SEMA, GSI-Erli, IBM, ASSTRIL, LADL (France), Un. Pisa, Un. Salerno, EDI-UTET-Paravia (Italy), ILTEC (Portugal)

- GRAAL

GSI-Erli, Aérospatiale, EDF, Rank Xerox, Renault (France), FIAT, IRST, Saritel (Italy), ILSP (Greece), ISSCO (Switzerl.), NOKIA, Lingsoft (Finland), ILTEC (Portugal)

Annex 5: Cocosda Members

Continent	Evaluation Recognition	Evaluation Synthesis	Corpus	Central Committee
Americas	<u>D. Pallett (US)</u> J. Baker (US) V. Zue(US)	M. Spiegel(US) K. Silverman(US) J. Olive(US)	G. Doddington(US) M. Picheny(US)	D. Pallett(US) M. Liberman(US)
Europe	J.L. Gauvain(F) R. Winski(UK) G. Castagneri(It)	<u>L. Pols(Netherl.)</u> B. Granström(Swed) C. Sorin(F)	J.M. Dolmazon(F) R. Moore(UK) I. Trancoso(Port)	J. Mariani(F) <u>A. Fourcin(UK)</u>
Japan	S. Furui	K. Shirai Y. Sagisaka	S. Itahashi S. Hayamizu	H. Fujisaki A. Kurematsu
China			J. Zhang	
Korea			M. Zhi	
Australia	M. Scordilis	J. Fletcher	<u>B. Millar</u>	B. Millar P. Dermody

Group coordinators are underlined.

Annex 6: Relator Industrial Steering Committee members

P. Alcouffe	Sema Group (France)
R. Billi	Cselt (Italie)
C. Blaesi	Bibliograpisches Institut & F.A. Brockhaus AG (Allemagne)
A. Castillo Helgado	Telefonica (Espagne)
C. Dugast	Laboratoires de recherche de Philips (Allemagne)
M. Filipe	TLP (Portugal)
A. Giannetti	Sogei (Italie)
G. Grimaldi	IBM-Europe
H. Höge	Siemens (Allemagne)
N. Kalfon	Grupo CL Servicios Linguisticos (Espagne)
H. Lehmann	IBM-Allemagne
G. Manos	Intrasoft (Grèce)
J. Massot	Matra Hachette (France)
R. Meyer	Cap Debis (Allemagne)
B. Normier	GSI Erli (France)
D. Pearce	GEC (GB)
J. Peckham	Vocalis (GB)
P. Procter	Cambridge University Press (GB)
A. Riccio	Alcatel face (Italie)
P. Rosenbeck	Jydsk (Danemark)
C. Roulin	BJL (Belgique)
T. Schneider	Sietec (Allemagne)
B. Seite	Eurolang (France)

Annex 7 : Business Plan for selling Spoken Language Resources

Year 1:

- Groningen (already distributed by Elsnet)	5	3	3	3	14
- Bref 80 (already distributed by Limsi)	20	20	0	0	40
- Bref / Polyglot (almost distributed by Elsnet)	20	20	0	0	40
- SAM Eurom 0 (to be distributed by Elsnet - 8 languages)	60	60	60	60	240
- SAM Eurom 1 (to be distributed by Elsnet - 11 languages)	60	60	60	60	240
- Maptask (also distributed by LDC)	20	20	20	20	80
- WJSCAM0 (also distributed by LDC)	30	20	10	10	70
- Phondat (already distributed in Germany)	20	10	10	10	50
Total #	235	213	163	163	774

Total sale public (KEcu)	64	62	56	51	233
Total sale industr. (KEcu)	120	92	36	57	305
Total sale (KEcu)	184	154	92	108	538

Year 2:

- ACCOR (to be distributed by Elsnet (English))	0	20	10	5	35
- TED (to be distributed by Elsnet)	0	15	15	15	45
- SCRIBE (already distributed in UK)	0	30	20	10	60
- Phondat-Verbmobil (to be also distributed in Germany)	0	30	20	10	60
- Mlap Speechdat (6 languages): "industrial price"	0	60	60	60	180
Total #	0	155	125	100	380

Total sale public (KEcu)	0	202	228	199	629
Total sale industr. (KEcu)	0	300	148	225	673
Total sale (KEcu)	0	502	376	424	1302

Year 3:

- Full Bref (developed for Francil)	0	0	30	30	60
- LDC DB (under cooperation agreement)	0	0	30	30	60
- Francil Dialog	0	0	20	10	30
- Francil Telephone DB	0	0	30	20	50
Total #	0	0	110	90	200

Total sale public (KEcu)	0	0	38	28	66
Total sale industr. (KEcu)	0	0	24	32	56
Total sale (KEcu)	0	0	62	60	122

Year 4: same as Year 3

(Babel, Polyglot etc)

Total #	0	0	0	90	90
----------------	----------	----------	----------	-----------	-----------

Total sale (KEcu)	0	0	0	60	90
--------------------------	----------	----------	----------	-----------	-----------

Grand total # sales for Speech:	235	368	398	443	1444
--	------------	------------	------------	------------	-------------

Total sale (KEcu) 1995 data	184	154	92	108	538
Total sale (KEcu) 1996 data	0	502	376	424	1302
Total sale (KEcu) 1997 data	0	0	62	60	122
Total sale (KEcu) 1998 data	0	0	0	60	60
Grand total amount sales for Speech (KEcu):	184	656	530	652	2022

Prices:

Member/subscriber Price for "public" data: 400 Ecu public / 1,600 Ecu industrial

Member/subscriber price for "industrial" data (Speechdat type): 4,000 public / 16,000 industrial
Non-member/subscriber price: + 50% (no sale foreseen here in that category)

Annex 8: Presentation of the LDC; subscription modalities and sale of data (1995)

The Linguistic Data Consortium (LDC), a nonprofit membership organization affiliated with the University of Pennsylvania, will add about 20 new releases to its 48 existing speech, text, and lexical databases during the current 1995 membership year. The new releases will feature text corpora in six languages, French-English parallel texts, a major telephone speech corpus, and new additions to the existing ARPA speech recognition and spoken language understanding series. Lexicons and large speech corpora in several languages are also in development and scheduled for release in six to nine months.

Consortium membership is annual, with the membership year (MY) running from September to August. Each LDC corpus is identified by the MY of its release, and the annual membership fee purchases a permanent paid-up license to that MY's releases, except that some corpora, owned by others and distributed by LDC, may require a separate user agreement and/or charges.

Members receive one copy of each requested LDC corpus free, and extra copies at a small charge. Nonmember prices are shown in the tables below. Items marked "MO" are for members only, due to restrictions by the copyright owners.

Detailed information about the LDC and a catalog describing its holdings are available via ftp or the World Wide Web (see below).

PRICES AND CONDITIONS OF PURCHASE

The following are the procedures and conditions for obtaining corpora from the LDC:

For LDC Members:

Membership fees for commercial organizations are \$20,000 per year; fees for non-profit organizations and government agencies are \$2,000 per year. Commercial members receive commercial rights to all resources, except where restricted by the original copyright holders. Notices are mailed to all members when new data sets are available. When corpora are re-issued in revised, enhanced, or supplemented form, unless the reason is defective materials, they will be distributed to all those whose LDC membership is current at the time of re-issue.

For Nonmembers:

With the exception of items marked "Members Only" (MO), nonmembers may purchase single copies of most listed items. Prices are set by the LDC from time to time, and normally include a permanent "research-only" license (i.e., no commercial use). Payment may be made by check drawn from a bank with branches in the United States or payment may be wired to: Mellon Bank East, ABA No. 03100003, Philadelphia, PA, for credit to The Trustees of the University of Pennsylvania, Account No 2945020, Attn: Sarah Parnum 215-898-0464.

Prices are subject to change; the prices above are effective until 1 June 1995. Nonmembers must purchase a minimum of \$200 in databases, and add a shipping charge for each order: \$30 US and Canada \$50 overseas.

FOR MORE INFORMATION, including membership forms and catalogs:

LDC is at ftp.cis.upenn.edu under /pub/ldc. When accessing by ftp, use "anonymous" as your userid, and your email address for password. The LDC's World Wide Web Home Page holds the LDC catalog and the "README" files from most of the databases. It can be accessed at URL:

ftp://ftp.cis.upenn.edu/pub/ldc_www/hpage.html

Annex 9: Data available at LDC (1995)

PLANNED 1995 RELEASES (TENTATIVE)

Non member Price	# of Disks	Title	LDC Catalog No
\$2500	1	KING Speaker Verification	LDC95S22
5000	2	Hansard French/English	May 1995
MO	3	CSR-III Speech: Dev and Eval Data	LDC9523
MO	4	CSR-III Text: Language Models	LDC95T24
2000	2	LATINO-40 Spanish Read News Corpus	April 1995
2000	6	*WSJCAM0: Cambridge Read News Corpus	LDC95S22
5000	5	PHONEBOOK: NYNEX Isolated Words	April 1995
2500	5	TRAINS spoken dialogs corpus	May 1995
2000	6	Corpus of Spoken American English-1	July 1995
2000	1	TIPSTER Volume 4	Spring 1995
2500	1	Treebank-2	March 1995
MO	1	Spanish News Text Collection	April 1995
MO	2	North American Business News Text	May 1995
MO	1	Japanese Business News Text	June 1995
2500	1	Mandarin News Text	May 1995
MO	1	*French Newspaper Text	August 1995
MO	1	North American Newspaper Text	August 1995
500	1	Speech Collection Interface SW	June 1995

PLANNED 1996 RELEASES (TENTATIVE)

Non member Price	# of Disks	Title	LDC Catalog No
TBA	14	JEIDA Japanese Speech Data	Summer 1996
TBA	12	Corpus of Spoken Amer English-2,3	1996
TBA	1	Mandarin Lexicon	Fall 1995
TBA	1	Spanish Lexicon	Fall 1995
TBA	1	Japanese Lexicon	Fall 1995
TBA	1	English Language International News	Fall 1995
TBA	3	Legal Text (500 M words)	Winter 1996
TBA	6	POLYPHONE-II (American Spanish)	Fall 1995
TBA	2	Mandarin Telephone Speech	Winter 1996
TBA	2	Japanese Telephone Speech	Winter 1996
TBA	2	Spanish Telephone Speech	Winter 1996
TBA	6	CALLFRIEND Language ID Corpus	Winter 1996
TBA	15	SWITCHBOARD (Revised)	TBA

1993 RELEASES

Non member Price	# of Disks	Title	LDC Catalog No
\$ 100	1	TIMIT	LDC93S1
250	2	NTIMIT	LDC93S2
750	6	Resource Management Complete	LDC93S3A
1000	6	ATIS0 Complete Set	LDC93S4A
2000	4	ATIS2	LDC93S5
2000	15	CSR-I (WSJ0) Complete	LDC93S6A
10000	28	SWITCHBOARD	LDC93S7
1000	1	SWITCHBOARD Credit Card	LDC93S8
125	1	TI 46-Word	LDC93S9
250	3	TIDIGITS	LDC93S10
250	1	Road Rally	LDC93S11
200	8	*HCRC Map Task Corpus	LDC93S12
25	1	ACL/DCI	LDC93T1
1000	1	TIPSTER Volume 1	LDC93T3-1.1
1000	1	TIPSTER Volume 2	LDC93T3-2.1
1000	1	TIPSTER Volume 3	LDC93T3-3.1

1994 RELEASES

Non member Price	# of Disks	Title	LDC Catalog No
10000	34	CSR-II (WSJ1) Complete	LDC94S13A
5000	19	CSR-II (WSJ1) Sennheiser	LDC94S13B
5000	20	CSR-II (WSJ1) Other	LDC94S13C
2500	8	Air Traffic Control	LDC94S14
2500	2	SPIDRE	LDC94S15
1000	1	YOHO Speaker Verification	LDC94S16
200	1	OGI Multilanguage Corpus	LDC94S17
100	1	OGI Spelled & Spoken Word	LDC94S18
5000	3	ATIS3	LDC94S19
2500	9	*BRAMSHILL	LDC94S20
10000	7	MACROPHONE (American English)	LDC94S21
5000	3	UN Parallel Text (Complete)	LDC94T4A
2500	1	UN Parallel Text (English)	LDC94T4B-1
2500	1	UN Parallel Text (French)	LDC94T4B-2
2500	1	UN Parallel Text (Spanish)	LDC94T4B-3.1
35	1	*ECI Multilingual Text	LDC94T5
150	1	*CELEX Lexical Database	LDC94L1
10000	1	COMLEX English Syntax Lexicon, Version 0	LDC94L2
10000	1	COMLEX Pronouncing Dictionary, Version 0	LDC94L3

*: *European corpus*