The ELRA Newsletter

October-December 2001

Vol.6 n.4

Contents

Signed articles represent the views of their authors and do not necessarily reflect the position of the Editors, or the official policy of the ELRA Board/ELDA staff.

Editor in Chief:
Khalid Choukri

Editors:
Khalid Choukri
Valérie Mapelli
Magali Duclaux-Jeanmaire

Layout:
Magali Duclaux-Jeanmaire

Contributors:
Catia Cucchiarini
Walter Daelemans
Helmer Strik
Pierre Zweigenbaum

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
55-57, rue Brillat Savarin
75013 Paris - France
Tel: (33) 1 43 13 33 33
Fax: (33) 1 43 13 33 30
E-mail: choukri@elda.fr or
WWW: http://www.elda.fr

Dear Members,

This year is coming to an end, and the activities shortly presented in this last issue of the ELRA newsletter for 2001 predict a good start for year 2002: on the one hand, the LREC (Language Resources and Evaluation) conference, a major milestone in the field of language engineering and for ELRA, will be organised in Spring 2002, in Las Palmas; on the other hand, ELRA & ELDA participate in several long term projects at French, European and international levels.

ELRA & ELDA have officially started the evaluation activity, which consists of evaluating Natural Language Processing technologies, systems, products, applications, tools, etc. We had tackled this activity some time ago, by supplying the language resources appropriate for testing and evaluation. ELRA & ELDA are now getting more involved in this area, and the tasks are being extended to other evaluation aspects, e.g. by fully participating in evaluation tasks for systems, tools, applications ,etc.

In order to stress this aspect; ELDA will modify the developed form of its name from ***E****uropean* ***L****anguage resources* ***D****istribution* ***A****gency* to ***E****valuation and* ***L****anguage resources* ***D****istribution* ***A****gency*.

The Euromap Language Technologies project brought together at the beginning of October the 'old' and 'new' participants in the project at the Ministère de la Recherche in Paris. One of the topics on the agenda was to discuss the public relations strategies and methods, e.g. the Euromap LT newsletter, the organisation of national and international events to promote Human Language Technologies (HLT), the dissemination of press releases, the publication of articles in national magazines and newspapers, the creation of a Euromap Web site, etc. The first steps have been achieved by ELDA, who has published on its Web site a Euromap page and has decided to localise the Euromap LT newsletter, which proves to be a success, both in its English and in its French version.

The partners of the Speecon project and some potential external partners met together in The Netherlands to amend the specifications, and discuss about the distribution media, the tests to be conducted on real-life applications, etc. Besides, ELDA is still proceeding to some recordings, in public places and car environments.

In the framework of the Network-DC project, ELRA/ELDA and its American counterpart LDC (Linguistic Data Consortium) are very much involved in the OLAC (Open Language Archives Community) initiative and fully participate in this project, which aims to set up a common Internet-based catalogue of language resources uniting the holdings of both ELRA/ELDA and LDC, along with several other resource providers. The other task, related to the collection of Arabic Broadcast news, initiated by ELDA, has also started in Paris with Radio-Orient, and 10 hours of Voice of America broadcast news have been selected by LDC.

The Mediterranean Language Technology Centre (MLTC) is now officially established in Morocco. A technical director has been recruited and will be in charge of the activities related to the OrienTel languages and of the tasks to be outsourced by ELRA/ELDA.

An ELRA Board meeting took place on 8th October: apart from the statistics on sales and membership, or the activities and the follow-up of the last strategic meeting, the process for the set up of an international Co-ordination Committee for the written language similar to Cocosda was presented, as well as the various activities related to the validation of resources, notably the set up of validation centres for written resources.

As for the LREC 2002 conference, 460 papers and 26 workshops proposals have been submitted to the LREC Programme Committee, whose members met on 2nd December in Paris. A reminder for the dates and the location is available at the end of this letter.

The content of this last issue for 2001 includes two articles in its first part, and the new resources added to the catalogue in a second part. The first article, entitled "Strengthening the Dutch Human Language Technology Infrastructure", was written by Catia Cucchiarini, Helmer Strik (both from The Netherlands) and Walter Daelemans (from Belgium). It gives an overview of the means and resources which are deployed for the development of language engineering tools and systems that integrate the Dutch language. The second article is entitled "Resources for the Medical Domain: Terminologies, Lexicons and Corpora.", and was written by Pierre Zweigenbaum, who presents in his paper how language resources can be used for the development of tools and systems dedicated to information processing in the medical field where the needs are constantly increasing.

Last but not least, listed below you can find the new resources distributed by ELDA. Their descriptions are given from page 12 to 15:

W0015: “Le Monde” Text Corpus (year 2000); W0029: Amaryllis Corpus; S0114: Strange Corpus 10 - SC10 ('Accents II'); S0115: American English SpeechDat-Car; S0116: Italian SpeechDat(II) MDB-250; S0117: Italian SpeechDat(II) FDB-3000; S0118: Greek SpeechDat(II) FDB-5000.

On the last page of the newsletter, you will find a membership form, that you can complete and return to the ELRA secretariat

Your comments and suggestions are welcomed.

The ELRA & ELDA team wish you a Merry Christmas and a Happy New Year.

Antonio Zampolli, President

Khalid Choukri, CEO

Strengthening the Dutch Human Language Technology Infrastructure

Catia Cucchiarini, Walter Daelemans and Helmer Strik

1. Introduction

The growing importance of Information and Communication Technology (ICT) in our society has emphasised the need for Human Language Technologies (HLT), since these make it possible for people to use natural language in their communication with computers. Preferably, this language should be the user's mother tongue, since this is the only way to guarantee that all citizens can fully participate in the information society. In order to develop HLT applications that allow people to use their native language in their interactions with computers, a digital language infrastructure is required for each language. By digital language infrastructure we mean all basic software tools, language and speech data, corpora and lexicons that are necessary for conducting research and developing applications in the field of HLT. Since the costs of developing HLT resources are high, it is important that all parties involved, both in industry and academia, co-operate so as to maximise the outcome of efforts in the field of HLT. This particularly applies to languages that are commercially less interesting than English, such as Dutch.

The last few years have witnessed a growing awareness of the importance of such a digital language infrastructure, not only in the United States and in Asia, but also in Europe. This is evident from the various initiatives that have been taken at European level, such as the creation of ELRA, the organisation of the LREC conferences, and the various projects funded by the European Commission, e.g. SPEECHDAT, PAROLE, SIMPLE, CLASS, EAGLES, HOPE, ISLE, to name but a few. Moreover, several projects have recently been launched by the National Authorities (Ministries or their Departments) in various European countries with the specific aim of strengthening the digital language infrastructure. To conduct projects of this kind, a dialogue has to be established between the parties involved, i.e. industry, academia and policy institutions. Such a dialogue is not always easy to establish, often because the various parties have conflicting interests. Discrepancies may exist not only between industry and universities, but also between the various research groups within industry and academia. From the contacts we have had with our European colleagues, it appears that it is just these kinds of problems that have hampered the emergence and the organisation of other countries' national projects aimed at providing or improving HLT resources for their respective languages.

In this paper we report on one such initiative that was taken for the Dutch language by the *Dutch Language Union* (*Nederlandse Taalunie* - abbreviated *NTU*): the Dutch Human Language Technologies platform. We hope that the experiences we have had in the last two years in setting up these activities may be useful to others who are now beginning with this kind of work.

2. The Dutch Language Union (NTU) and Human Language Technologies

The plan to set up a Dutch HLT platform was launched by the NTU. This is an intergovernmental organisation established in 1980 on the basis of the Language Union Treaty between Belgium and the Netherlands, which has the mission of dealing with all issues related to strengthening the position of the Dutch language (see also www.taalunie.org). The NTU enables Flanders and the Netherlands to speak with a single voice in the international arena. The Committee of Ministers, composed of the Flemish and Dutch ministers of Education and Culture, is responsible for the policy of the NTU. In establishing its current long-term policy plan (1998 - 2002), the NTU has given full consideration to the rapid developments in the field of ICT that are going to have a major impact on language issues. The governments of the Netherlands and Flanders appreciate the growing importance of HLT as a specific part of information technology. Keeping up with the technological developments in this field implies major investments and the commitment of those involved, notably the policy makers at the national and European level, the knowledge infrastructure and the business community. Co-operation among all these actors is of utmost importance, and given the size of the Dutch language area, this co-operation needs to be expanded to a cross-border Flemish-Dutch level. Building on this awareness, two large HLT projects, that were initiated over the last years, not only have a Flemish-Dutch character but also try to combine expertise from the research community as well as of the business community.

The Spoken Dutch Corpus project is a five-year project, which aims at the compilation and annotation of a 10-million-word corpus of contemporary standard Dutch as spoken in the Netherlands and Flanders (see also Oostdijk, 2000). The project is funded jointly by the Dutch and Flemish governments. Project activities are co-ordinated from two sites: one in Flanders and one in the Netherlands. The copyright to the Spoken Dutch Corpus is owned by the NTU who will be responsible for the exploitation of the results.

NL-Translex is a project which aims at the development of machine translation modules for the language pairs Dutch - English/French and English/French- Dutch (see also Cucchiarini, 2001; Goetschalckx, Cucchiarini, and Van Hoorde, 2001) . The development of these components takes place within the framework of MLIS. The project is funded jointly by the European Commission, the Dutch Language Union, the Dutch Ministry of Education, Culture and Science, the Dutch Ministry of Economic Affairs, the Flemish Institute for the Promotion of Scientific and Technological Research in Industry, and Systran, which is the technology provider. The components to be developed are intended for use by the translation services of official bodies of the EU Member States and by the translation services of the European Commission.

In the project preparation of the Spoken Dutch Corpus as well as of NL-Translex much time was spent in finding the appropriate responsible (funding) bodies as it was not clear who was responsible for the construction of a digital language infrastructure for Dutch. This observation was confirmed in several surveys that were conducted over the last few years. The market research carried out in the Netherlands and in Flanders in the framework of EUROMAP Language Technologies and the research commissioned by the NTU into the position of Dutch in Language and Speech Technology (report Bouma and Schuurman, 1998) pointed out that the fragmentation of responsibilities made it difficult to conduct a coherent policy and meant that the field lacked transparency for interested parties. In order to create more transparency and to give shape to the co-operation in the field of HLT, the NTU took the initiative to set up a Dutch-Flemish platform to support the Dutch language in HLT.

3. The Dutch Human Language Technologies Platform

The main purpose of the Dutch HLT platform is to further development of an adequate digital language infrastructure for Dutch so that the applications can be developed which can guarantee that the citizens in Holland and Flanders can use their own language in their communication within the information society and that the Dutch language area remains a full player in a multi-lingual Europe.

More specifically, the HLT platform has the following objectives:

- to strengthen the position of the Dutch language in HLT developments, so that the Dutch speakers can fully participate in the information society;

- to establish the proper conditions for a successful management and maintenance of basic HLT resources developed through governmental funding;

- to stimulate co-operation between academia and industry in the field of HLT;

- to contribute to the realisation of European co-operation in HLT-relevant areas;

- to establish a network that brings together demand and supply of knowledge, products and services.

In addition to the NTU, the following Flemish and Dutch partners are involved in the HLT platform:

- the Ministry of the Flemish Community;

- the Flemish Institute for the Promotion of Scientific-technological Research in Industry;

- the Fund for Scientific Research - Flanders;

- the Dutch Ministry of Education, Culture and Sciences;

- the Dutch Ministry of Economic Affairs;

- the Netherlands Organisation for Scientific Research (NWO);

- Senter (an agency of the Dutch Ministry of Economic Affairs).

All these organisations have their own aims and responsibilities and approach HLT accordingly. Together they provide a good coverage of the various perspectives from which HLT policy can be approached.

The rationale behind the Dutch HLT platform was not to create a new structure, but rather to co-ordinate the activities of existing structures. The platform is a flexible framework within which the various partners adjust their respective HLT agendas to each other's and decide whether to place new subjects on a common agenda. Initially, the Dutch HLT platform was set up for a period of five years (1999-2004).

Even if the Netherlands and Flanders co-operate in funding the development of basic language resources, the investments for the different partners involved remain substantial. This absolutely requires that efforts be cumulative and not duplicated, that insight be provided into the resources that are needed for a language in general and for Dutch in particular and that a plan be drawn up for the development of the resources that are totally lacking or insufficiently available for Dutch. Furthermore, attention should be paid to such matters as evaluation of resources and project results, standardisation, maintenance, distribution etc. In other words, it is necessary to create the preconditions to maximise the outcome of efforts in the field of HLT. To this end, an action plan for Dutch in language and speech technology has been defined, which is funded jointly by the different partners in the HLT platform. The activities described in this action plan are organized in four action lines:

Action line A: performing a 'market place' function

The main goals of this action line are to encourage co-operation between the parties involved (industry, academia and policy institutions), to raise awareness and give publicity to the results of HLT research so as to stimulate market takeup of these results.

Action line B: strengthening the digital language infrastructure

The aims of action line B are to define what the so-called BLARK (Basic LAnguage Resources Kit) for Dutch should contain and to carry out a survey to determine what is needed to complete this BLARK and what costs are associated with the development of the material needed. These efforts should result in a priority list with cost estimates which can serve as a policy guideline.

Action line C: working out standards and evaluation criteria

This action line is aimed at drawing up a set of standards and criteria for the evaluation of the basic materials contained in the BLARK and for the assessment of project results.

Action line D: developing a management, maintenance and distribution plan

The purpose of this action line is to define a blueprint for management (including intellectual property rights), maintenance, and distribution of HLT resources.

In this paper we will focus on action lines B and C.

4. Action lines B and C: survey, evaluation and directions for future development

As explained in section 2, the purpose of action line B is to define the BLARK for Dutch and to determine what should be developed on the basis of a detailed analysis of the needs for HLT resources in the short and medium term, in comparison with the BLARK definition and the present situation.

However, it is not sufficient to acknowledge the existence of a given resource, be it a piece of language data or a tool: all HLT resources, to be really useful, have to meet requirements of formal and content quality, availability (free of rights or under certain conditions), multi-functionality and re-usability. It follows that the work to be carried out for action line B is inextricably linked to the activities in action line C. Only on the basis of a qualitative evaluation is it possible to establish whether the resources that already exist are available and qualitatively satisfactory. This gives a clearer view of what can be included in the HLT infrastructure. The results of such an analysis will reveal which materials are suitable, unsuitable (for example not multifunctional or not available) or are only suitable after adaptation. This will provide a realistic view on the present state of affairs with respect to HLT resources. For the reasons mentioned above, it was soon decided that action lines B and C would be carried out in an integrated way.

In the following sections we provide more detailed information on action lines B and C. First we describe the structure that was set up to conduct the work planned in these two action lines. We then describe the tasks of the various participants. Subsequently, we present the instruments that were developed to carry out these activities and, finally, we present the results obtained so far.

4.1. Structure

4.1.1. Steering committee

The first step in organising the activities for action lines B and C was to set up a Flemish-Dutch steering committee. This committee is composed of experts from different disciplines in HLT and of representatives of language and research policy institutions such as NTU and NWO. The experts have been selected on the basis of their nationality and their expertise. More precisely, there are four experts from the Netherlands and four experts from Flanders. For each geographical area there are two experts on language technology and two experts on speech technology. This composition guarantees that all parties involved have a representative that will protect their interests and that will provide reliable information on the topics at issue.

The steering committee has the followings tasks:

- to draw up a plan of the activities that should be carried out to achieve the goals of action lines B and C;

- to develop an initial framework that will be used for surveying the current state of Dutch HLT resources;

- to select and hire field researchers who will carry out the actual field survey (see following section);

- to supervise the field survey of Dutch HLT resources;

- to establish a set of standards and evaluation criteria for HLT resources;

- to define the so-called BLARK (Basic LAnguage Resources Kit) for Dutch;

- to draw up a list of what is needed to complete the BLARK and what costs are associated with the development of the material needed.

The framework to be used in the field survey will be presented in the following section.

4.1.2. Field researchers

Four field researchers have been appointed by the steering committee, two for language technology and two for speech technology. These researchers have the following tasks:

- to further refine the framework that will be used for surveying the current state of Dutch HLT resources;

- to develop specific instruments for the field survey (tables and questionnaires);

- to collect information on HLT evaluation instruments;

- to conduct the field survey;

- to write a report.

4.2. Survey framework

In order to carry out a thorough survey of the current state of Dutch HLT resources, adequate instruments are needed which guarantee, as much as possible, that the survey is complete, unbiased and uniform. The HLT experts in the steering committee worked out an initial framework that was further refined by the field researchers. In setting up this framework, the experts have used the usual three components: 1. applications, 2. modules and 3. data.

4.2.1. Applications

In this framework, the term application refers to a class of applications rather than to a specific application or product. This is done to obtain a framework that is general enough to capture all sorts of possible applications. The applications distinguished are:

- *CALL* (*Computer Assisted Language Learning*).

- *Access control:*

Applications in which physical characteristics such as speech signals are used for speaker verification or identification to provide access to systems, buildings etc.

- *Speech input:*

Applications in which speech input is analysed and converted into text. This category also includes applications such as command and control, dictation, and automatic transcription.

- *Speech output:*

Applications in which text is converted into speech, such as spoken e-mail, spoken dictionaries and aids for the blind.

- *Language and speech interfaces:*

Spoken dialogue systems that constitute a natural interface to databases, expert systems, information systems and virtual reality applications in which speech interaction plays a part.

- *Document production:*

All applications concerning text production, from spelling, grammar and style checking up to text generation.

- *Information access:*

Applications in which text and speech analysis play a part in information localisation and knowledge extraction, information retrieval, text mining, document routing, filtering and classification, question answering etc.

- *Machine translation:*

Translation aids, translation memories, machine translation.

4.2.2. Modules (semi-products)

Under modules, or semi-products, we understand the basic software components of HLT applications. In general, these components do not have much commercial value as such, but they are essential in the HLT infrastructure. The list of modules identified so far is given below:

- Rule-based synthesis;
- Diphone synthesis;
- Unit selection;
- Sentence boundary detection ;
- Grapheme-phoneme conversion;
- Complete speech synthesis;
- Complete speech recognition;
- Token detection;
- Lemmatising;
- Morphological analysis;
- Morphological synthesis;
- Part of speech tagging;
- Constituent recognition;
- Shallow parsing;
- Named entity recognition;
- Parsers and grammars;
- Prosody prediction;
- Referent resolution;
- Word meaning disambiguation;
- Semantic analysis;
- Pragmatic analysis;
- Text generation;
- Language-pair dependent translation modules.

4.2.3. Data

In this case, the term data refers to sets of language data and descriptions in machine readable form, to be used in building, improving or evaluating natural language and speech processing systems. Examples of data are written and spoken corpora, lexical databases and terminology lists. In our scheme the following data types have been distinguished:

- *Monolingual lexicons:*

Lexicons containing orthographic, phonetic, phonological, morphological, syntactic, semantic and pragmatic knowledge about lexical entities (morphemes, word forms, collocation and special expressions).

- *Multilingual lexicons:*

Monolingual lexicons with translations of the lexical entities.

- *Thesauri:*

Lexicons with semantic and associative relations among words.

- *Annotated text corpora:*

Large (10M+) text databases with annotation tiers for orthography, phonology, morphology, syntax, semantics and pragmatics. These data are especially important for training the various modules.

- *Non-annotated text corpora:*

Large (100M+) text databases without annotation tiers, which contain some information about the origin of the texts and, possibly, the typographic structure. These corpora are used for unsupervised training.

- *Speech corpora:*

(10M+) databases with, at least, orthographically annotated speech.

- *Multingual corpora:*

Databases that contain speech from Dutch and other languages.

- *Multimodal corpora:*

Databases that contain speech and data from other modalities

- *Multimedia speech corpora:*

Databases that contain speech from radio and TV but also information from other media (e.g. texts and figures from WWW, papers and journals, etc.).

4.2.4. Matrices

On the basis of the relationships between the three components mentioned above, applications, modules and data, three matrices were designed that address three different topics in the HLT infrastructure:

- *Relevance of modules for applications:*

This matrix shows which modules are required for the various applications.

- *Relevance of data for modules:*

This matrix shows which data are required for the various modules.

- *Availability of data and modules:*

This matrix indicates which data and modules are really available in the sense that they have an acceptable quality level.

4.3. Survey results

4.3.1. BLARK

On the basis of matrices 1 and 2, a BLARK for language technology and one for speech technology can de derived. These are shown below:

BLARK for language technology

Modules

- Robust modular text preprocessing;
- Morphological analysis and morphosyntactic disambiguation / unknown words;
- Robust syntactic analysis;
- Aspects of semantic analysis (word meaning and reference).

Data

- Monolingual lexicon;
- Annotated corpus of written Dutch;
- Benchmarks for evaluation.

BLARK for speech technology

Modules

- Automatic speech recognition (module);
- Speech synthesis system (module);
- Tools for annotation of speech corpora;
- Confidence measures and utterance verification;
- Identification (speaker, language, dialect);
- Evaluation of speech technology tools and applications.

Data

- Monolingual speech corpora for specific applications;
- Multilingual speech corpora;
- Multimodal/media speech corpora;
- Richly annotated speech corpora;
- Pronunciation lexicons.

4.3.2. List of priorities

By analysing the availability of modules and data, priority can be assigned to the development of those parts of the BLARK that are known to be crucial and appear to be missing. The general idea is that those components and data that are relevant for many applications and turn out to be unavailable or of low quality should be developed first. On the basis of matrix 3, such a list of priority was drawn up and was subsequently submitted to representatives from the whole HLT field. The results of this consultation are to be discussed in public during a supranational seminar to be held in The Hague on 15th November (for further information, the reader is referred to http://www.taalunieversum.org/tst/).

5. Concluding remarks

In this paper we have reported on the activities that were carried out in the two years that the Dutch HLT platform has been active. It should be noted that much effort was spent in setting up the whole platform structure, i.e. in finding the representatives of the appropriate responsible bodies and expertise centres. Owing to the fragmentation of responsibilities, it was difficult in the past to conduct a coherent HLT policy. We hope that the HLT platform will contribute to creating more transparency in this respect.

Up to now our experiences have been positive across the board. It turned out that experts from different disciplines and different countries managed to work together and could reach an agreement on a number of important matters. We can only hope that this trend will continue, since there is still much work to be done.

6. References

Bouma, G. and Schuurman, I. (1998) *De positie van het Nederlands in Taal- en Spraaktechnologie*. Report for the Dutch Language Union.

Cucchiarini, C. (2001) *The Dutch Connection. A European Machine Translation project for the Dutch Language*, in *Language International*, Vol. 13 No. 4, 43-46.

Goetschalckx, J. Cucchiarini, C. and Van Hoorde, J. (2001) *Machine Translation for Dutch: the NL-Translex Project Why Machine Translation?*, in *Proceedings of the International Colloquium Trends in Special Language & Language Technology*, R. Temmerman & M. Lutjeharms (eds.), 261-280.

Oostdijk, N. (2000) *The Spoken Dutch Corpus. Overview and first Evaluation*, in *Proceedings LREC 2000*, Athens, Greece.

7. Acknowledgements

We are indebted to the other members of the steering committee and to the field researchers for their invaluable contribution to the work presented in this paper: J. Beeken, D, Binnenpoorte, G. Bouma, E. Dewallef, E. D'Halleweyn, F. De Vriend, A. Dijkstra, A. Dirksen, D. Heylen, M. Kempen, F. de Jong, J-P. Martens, A. Nijholt, J. Sturm, D. Van Compernolle, F. Van Eynde, R. Veldhuis.

Dutch Language Union
Lange Voorhout 19
P. O. Box 10595
2501 HN The Hague
Tel.: +31 70 3469548
Fax: +31 70 3659818

Catia Cucchiarini
A2RT, Department of Language & Speech
University of Nijmegen
P.O. Box 9103
6500 HD Nijmegen (The Netherlands)
Tel.: +31 24 3615785/Fax: +31 24 3612907
Email: catia@let.kun.nl
Web site: http://zap.to/catia

Helmer Strik
A2RT, Department of Language & Speech
University of Nijmegen
P.O. Box 9103
6500 HD Nijmegen (The Netherlands)
Tel.: +31 24 3616104/Fax: +31 24 3612907
Email: strik@let.kun.nl
Web site: http://zap.to.helmer

Walter Daelemans
CNTS Language Technology Group
University of Antwerp (UIA)
Universiteitsplein 1, Building A
B-2610 Antwerpen (Belgium)
Tel.: +32 3 8202766/Fax: +32 3 8202762
Email: daelemans@uia.ua.ac.be

Resources for the Medical Domain: Terminologies, Lexicons and Corpora

Pierre Zweigenbaum

1. Introduction

A wealth of textual documents of all sorts can be found in the medical domain: reports in hospital patient records, scientific articles, textbooks, practice guidelines are but a few examples. Besides, medicine is characterised by a vast amount of knowledge, which is mirrored by a particularly large lexicon and terminology.

The need for processing information and knowledge conveyed by medical texts motivates work in medical language processing. This is the research theme of our team. Whereas terminologies have long been established for this domain, lexicons and corpora are rarer; we have therefore endeavoured to build some. We present here successively the main terminologies that are available for the medical domain, then our own work on the construction of lexicons and corpora. Complementary information can be found in the cited references, in particular [1,2,3,4].

2. Terminologies

Collecting medical information can have three main objectives. Each objective corresponds to a type of terminology. We briefly present each one in turn, providing examples and information about their availability.

The first objective is the collection of a relatively limited set of data elements for statistical purposes. For instance, for a given patient, the main diagnosis and if relevant its associated diagnoses. This is the case for the evaluation of the activity of hospital departments (e.g., the French "Programme for the Medicalization of the Information System" or PMSI) and in epidemiological studies. Patients are categorised into non-overlapping classes, and the terminologies used are structured as hierarchical classifications. The International Classification of Diseases (ICD-10) [5], which mainly gathers diagnoses, is the most widespread example of such a classification. Created one century ago, the ICD is the official terminology for the compilation of national statistics on mortality and morbidity in most European countries, e.g., for France's PMSI. ICD-10 is managed by the World Health Organisation (WHO) and exists for a large number of languages among which French. It contains about 11,000 "rubrics" (diagnostic terms), and is complemented with an index that specifies numerous equivalent or semantically close terms. The ICD-10 classification is sold by the WHO (http://www.who.ch/hst/icd-10/icd-10.htm).

The second objective is the description of more varied information items which are to be stored in a patient's record. As the information need is more open and the required level of detail may be higher, a finer precision and the coverage of a larger palette of notions are necessary. The *Systematised Nomenclature of Medicine* (SNOMED International) [6] is an instance of the terminologies designed with this aim. It compiles about 100,000 concepts (about 160,000 terms including synonyms) organised within eight orthogonal semantic axes: anatomy, morphological conditions, functional conditions, living beings, etc., complemented with two classifications (diagnoses and procedures) and a set of “qualifiers and relational terms”. These concepts can be combined to describe a given medical fact. This provides this nomenclature with an expressive power far superior to terminologies that must a priori enumerate all the concepts of the target field (so-called pre-coordinated terminologies). And indeed, in the studies led so far (e.g., [7]), SNOMED is the medical terminology which offers the best coverage of clinical data. SNOMED, created by Roger Côté's team in Canada, is managed and sold by the College of American Pathologists (CAP, http://www.snomed.org/). A subset of SNOMED, the *Microglossary for Pathology*, has been translated into French; it is sold by the MedSight company (http://www.medsight-info.com/en/Products/fsnomed.htm). It contains 12,555 terms. The French translation of the full SNOMED is ongoing. Translations are also ongoing or available in other languages; among other examples, the *Microglossary for Pathology* exists in Russian and Japanese and the full SNOMED has recently been translated to Spanish in Argentina.

The third objective is the indexing of scientific articles for bibliographic search purposes. The terminology built to answer this need is the *Medical Subject Headings* (MeSH) [8]. It was created by the US National Library of Medicine (NLM) and indexes its Medline online bibliographic database. The MeSH is also used to index an increasing number of catalogues of Internet medical sites (e.g., in France, the CISMeF directory, http://www.chu-rouen.fr/cismef/, and in Switzerland, the HON directory, http://www.hon.ch/). As in SNOMED, several semantic fields are involved. The MeSH aims at a lower precision though, but at a wider domain: e.g., it includes a "geographical" axis (countries, main cities) whereas SNOMED does not. The MeSH contains about 20,000 "concepts" expressed by about 40,000 different terms. It may be downloaded from the NLM web site (http://www.nlm.nih.gov/mesh/meshhome.html). The MeSH has been translated in several languages. Its French version was written by INSERM, which continues to cater for its annual update. The French MeSH contains about 28,000 terms and can be visited on the INSERM site .(http://dicdoc.kb.inserm.fr:2010/basismesh/mesh.html). For historical reasons, its terms are written in unaccented uppercase letters.

A survey, be it short, of the main medical terminologies cannot be concluded without mentioning the *Unified Medical Language System* (UMLS) [9]. Its "metathesaurus" compiles in 2001 the terms of more than 60 biomedical terminologies, amounting to an order of 800,000 concepts and 1,900,000 terms. It includes, among others, the English versions of the three above-mentioned terminologies, as well as translations of the MeSH, including the French version by INSERM. The UMLS may be freely obtained for research purposes after signing an agreement with the NLM (http://www.nlm.nih.gov.research/umls/).

3. Lexicons

The above-mentioned terminologies compile the concepts of a domain and their associated terms. To fulfil the needs of natural language processing, they must be complemented by lexicons that associate linguistic information with the relevant lexical entries.

As a matter of fact, so-called “general” electronic lexicons may contain numerous specialised terms. For instance, applying the LADL's DELAF lexicon [10] to ten patient discharge summaries in pneumology [2] obtained a coverage of 88% of the 5,896 word occurrences found in this corpus, including a number of specialised words (acidose, adénopathie, amylasémie, etc.). Among the remaining 12%, four main classes can be found: drug names, abbreviations, specialised words which for the most part are constructed by derivation or neo-classical compounding from known bases (hépatitique, hémodiafiltration), and proper names of patients, of doctors, of hospitals, etc. The same kind of experiment, applied to the terms of a medical terminology, shows a greater proportion of unkwnown words. Among the 16,132 terms of the French version of the 1992 MeSH thesaurus, 20% of the 29,604 word occurrences were unkwnown of the DELAF (36% of the 13,255 different word forms).

Among the four classes of specific words identified above, the class of drugs can be easily covered: drug databases, such as

Vidal (http://www.vidal.fr/) and Thériaque (http://www.theriaque.org/), contain an extensive list of their common denominations, commercial names and component substances. Some common abbreviations are listed in specialised terminologies or printed dictionaries. However, any useful word may be abbreviated in a text, and it would be vain to try and make an exhaustive a priori inventory of all possible abbreviations.

We have worked on the two remaining classes of specific words: morphological constructs and proper names. We sum up below the method we used to collect some derivations and compounds which can be found in the medical lexicon, based on the study of structured terminologies [3]. We then address the collection of proper names that occur in medical terms (eponyms) [11]. We have furthermore prepared, semi-automatically or manually, lexicons for the words that occur in several medical terminologies: ICD-10, SNOMED Microglossary for Anatomy, and several specialty thesauri [2].

3.1 Morphology

As we saw above, large-size medical terminologies exist in French as well as in other languages. In contrast, morphological knowledge bases are less common. The CELEX base contains general derivational knowledge for English, German and Dutch. The "Specialist Lexicon", which is a companion to the UMLS metathesaurus [12], contains tools and derivational knowledge for medical English. However, such knowledge bases are not publicly available for French.

We have set up a method to extract automatically from medical terminologies (SNOMED Microglossary and ICD-10) relational morphological knowledge: word pairs related by inflection (abdominalabdominale), derivation (abdomenabdominal) and compounding (adénomeadénofibrome) [3]. This method relies on the availability of a terminology that includes semantic relations between terms (synonymy, hierarchy, etc.). When two semantically related terms contain two words whose form is similar, there are chances that these words be morphologically related. For instance, the SNOMED nomenclature states that "sinusite, SAI" is a kind of "paranasal sinus disease, SAI" (SAI, or NOS, means *Sans Autre Indication*, or *Not Otherwise Specified*). We then hypothesise that the words "sinussinusite" are morphologically related. We also induce that a suffix substitution rule Eite is at work and can apply to other couples of words attested in the domain. When applied to these terminologies, this method generates very little noise (3 to 5%): nearly all the pairs that are found involve words that are actually morphologically related.

The method and the programs are independent of the processed language; we could apply them to English-language terminologies (full SNOMED, ICD-10) and compare the results to the inflectional and derivational knowledge contained in the UMLS Specialist Lexicon: coverage of the word pairs is over 88% with a precision over 90%. We have also worked on Russian (SNOMED Microglossary) with the same programs. Finally, derivational knowledge selected from the obtained data have been tested to perform query expansion in French for a specific information retrieval task [13], and increased by a few percents the results obtained for the queries of our test set (obtained from an actual log of Web queries).

3.2 Proper names

In this stream of work on medical terminologies, we have tried to collect the proper names that occur in their terms [11]. These proper names are used to construct a palette of eponym terms: diseases, signs and symptoms, tests, medical implements, diagnostic or therapeutic procedures, micro-organisms or body parts. The names are those of prominent scientists (Parkinson disease), places (Lassa fever), historical (Münchhausen syndrom) or mythological characters (Achilles' tendon). Their form ranges from a simple name, as in the previous examples, to compound names (Swan-Ganz catheter), coordinated names (Pierre Marie and Sainton disease) which may also be preceded by christian names and particles.

Since medical terminologies are numerous and large-sized, we tried to design a generic, customisable method to help collect proper names occurring in a terminology. This is a restricted version of the named entity recognition task: decide whether a word is a proper name, based on isolated terms rather than full sentences. The difficulty comes from the fact that case alone (starting with an uppercase letter) is not an absolute criterion. We therefore listed a series of complementary criteria which, when satisfied, tend to show that a word is a proper name. Each criterion is insufficient in isolation, but an appropriate combination of these criteria is a strong clue for the identification of proper names. These criteria are the following: starting with an uppercase letter in all its occurrences in the considered terminology; not being a symbol or an abbreviation (AIDS); not being a micro-organism genus name, which by convention is always capitalized (Hæmophilus influenzae); being invariant through several translations (Sarcome de Kaposi de la peau, Kaposi's sarcoma of skin); occurring in specified positions (N) in a lexico-syntactic pattern such as X de N (for instance, maladie de Parkinson). The combination of criteria that gives the best balance between precision and recall, on the terminologies studied, is specified as [("capitalised" and not "symbol" and not "micro-organism") and ("invariant"or "pattern")]. It reaches a joint score of 86% precision and 88% recall on ICD-10 (1,222 proper names) and 98% precision and 97% recall on the SNOMED Microglossary for Pathology (339 proper names).

4 Corpus

Medical language processing has focussed until recently on a few types, or genres, of textual documents. However, a much larger variety of document types are produced and read in diverse situations. The performance of NLP tools may substantially vary from one genre to another (see, e.g., [14,15]). Therefore, it would be very useful to make available a corpus of medical text displaying a large palette of genres so that one can tune, train or test NLP tools in this domain.

In reference corpora, the coverage of specialised domains does not always account for the diversity of textual genres in these domains. In a similar way, most studies on sublanguages have concentrated on domain specialisation, generally leaving aside the genre as an implicit choice. We therefore endeavored to build a corpus of medical texts that focuses on representing the main genres in use in that domain [4]. We have until now prepared the framework for designing this corpus: an inventory of the main genres of the domain, a set of dimensions for describing documents and a normalised encoding of this description (meta-information) and of their contents. We also implemented a demonstrator that encodes an initial text corpus documented and encoded according to these principles. This work takes place within the CLEF project (http://www.biomath.jussieu.fr/CLEF/), coordinated by Benoît Habert, which aims to build a diversified corpus of current French.

Several medical text corpora have been created in the past - let us mention, e.g., projects LECTICIEL [16] and MEDICOR [17] -, but none of them tried to cover and document a wide variety of text genres. Besides, whereas large document bases exist in hospitals (patient discharge summaries, test reports, etc.), issues of privacy and anonymisation do arise.

Building a complete list of medical text genres is probably a never ending task, since new situations may lead to the creation of new document types, and finer distinctions can always be made. Our goal is rather to identify the main types of medical texts which can be found in electronic form, and to characterise them according to a set of orthogonal dimensions.

We considered four main contexts of production or use of medical documents - some of these contexts may intersect. In the context of care, health care professionals produce information about a patient; this information is conveyed by reports (discharge, test, etc.) and letters to other physicians (in other hospital departments or outside the hospital). In a university context, teaching motivates documents produced by the faculty (books, syllabi, tests) and by students (student notes). In a research context, health care professionals read and write articles in various journals and conferences, at different levels of vulgarisation. We included there thesis manuscripts, as well as information exchange through discussion lists and newsgroups. Finally, various reference knowledge sources are used in the medical practice: dictionaries, encyclopedias and monographs (e.g., about drugs); practice guidelines and diagnostic or therapeutic protocols; consensus conferences; official documents (Bulletin officiel, Code of deontology, convention); and coding systems, such as the above-mentioned terminologies.

As can be seen, it may be difficult to avoid intersections between groups when clustering together these document types. It is therefore all the more interesting to characterise more finely these classes and their individual documents, by describing each of them according to a number of predetermined dimensions (or features). We have built on sets of dimensions proposed in the literature [18,19,20]. These dimensions belong to three groups: the bibliographic dimensions are the usual attributes that describe the origin of a document (author, etc.); the other external dimensions characterise the context of production or reception of the document (written, spoken; profiles of producer and of receptor; published or not; etc.); the internal dimensions can generally be determined from the text itself (language, size, level of language, technicity, etc.).

The specific domain of the text constitutes a particular dimension. Medicine is indeed the main domain of the texts involved here, but it is subdivided into numerous specialties; we relied on the list used by the Catalog of French language Medical Internet Sites (CISMeF, http://www.chu-rouen.fr/cismef/).

The corpus is encoded in XML, according to the Text Encoding Initiative Corpus Encoding Standard (TEI XCES, [21]). Each text includes a document header in which values are defined for the above dimensions. A corpus header similarly documents the whole corpus. This way, the corpus can be processed with standard XML tools. We wrote transformation stylesheets that allow to extract sub-corpora as needed.

To initialise the construction of the corpus, we selected a subset of the above types. The demonstrator contains 374 documents distributed among patient discharge summaries (4 different sites, cardiology and hematology), discharge letters, a chapter of a book on coronary angiographies and a consensus conference (post-operative pains), for a total of 143 Kwords. Many colleagues have kindly declared their intent to provide us with documents, which will help to scale up the corpus. For documents where patients are involved, we contacted the French Council for Informatics and Liberties (CNIL) which considered that these documents would be distributable once proper names and dates are hidden. This is the case of the above-mentioned reports and letters.

This corpus, once sufficiently completed, will be distributed. We hope it will prove useful for researchers and engineers to test and train NLP methods and tools, and for linguists to study the features of medical language.

Bibliography

[1] Zweigenbaum P. *Traitements automatiques de la terminologie médicale.* In Revue française de linguistique appliquée 2002. Submitted.

[2] Zweigenbaum P. *Des lexiques pour la terminologie médicale.* In Lingvisticae Investigationes 1998/1999;1/2(22):383-95.

[3] Grabar N and Zweigenbaum P. *Automatic acquisition of domain-specific morphological resources from thesauri*. In: Proceedings of RIAO 2000: Content-Based Multimedia Information Access, Paris, France. C.I.D., April 2000:765-84.

[4] Habert B, Grabar N, Jacquemart P, and Zweigenbaum P. *Building a text corpus for representing the variety of medical language*. In: Corpus Linguistics 2001, Lancaster. 2001.

[5] Organisation mondiale de la Santé, Genève. *Classification statistique internationale des maladies et des problèmes de santé connexes* - Dixième révision, 1993.

[6] Côté RA, Rothwell DJ, Palotay JL, Beckett RS, and Brochu L, eds. *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED* International. College of American Pathologists, Northfield, 1993.

[7] Chute CG, Cohn SP, Campbell KE, Oliver DE, and Campbell JR. *The content coverage of clinical classifications*. In Journal of the American Medical Informatics Association 1996;3(3):224-33. for the Computer-Based Patient Record Institute's Work Group on Codes and Structures.

[8] National Library of Medicine, Bethesda, Maryland. Medical Subject Headings, July 1986.

[9] Lindberg DAB, Humphreys BL, and McCray AT. *The Unified Medical Language System*. In Methods of Information in Medicine 1993;32(2):81-91.

[10] Silberztein M. *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*. Masson, Paris, 1993.

[11] Bodenreider O and Zweigenbaum P. *Stratégies d'identification de noms propres à partir de nomenclatures médicales parallèles*. In Traitement automatique des langues 2000;41(3):727-57.

[12] McCray AT, Srinivasan S, and Browne AC. *Lexical methods for managing variation in biomedical terminologies*. In: Proceedings of the 18th Annual SCAMC, Washington. Mc Graw Hill, 1994:235-9.

[13] Zweigenbaum P, Grabar N, and Darmoni S. *L'apport de connaissances morphologiques pour la projection de requêtes sur une terminologie normalisée*. In: Maurel D, ed, Proceedings of TALN 2001 (Traitement automatique des langues naturelles), Tours. ATALA, Université de Tours, July 2001:403-8.

[14] Friedman C. *Towards a comprehensive medical natural language processing system: Methods and issues*. In Journal of the American Medical Informatics Association 1997;4(suppl):595-9.

[15] Illouz G. *Méta-étiqueteur adaptatif : vers une utilisation pragmatique des ressources linguistiques*. In Amsili P, ed, Proceedings of TALN 1999 (Traitement automatique des langues naturelles), Cargèse. ATALA, July 1999:185-94.

[16] Lehmann D, de Margerie C, and Pelfrêne A. *Lecticiel - rétrospective 1992-1995*. Technical report, CREDIF - ENS de Fontenay/Saint-Cloud, Saint-Cloud, 1995.

[17] Vihla M. Medicor: *A corpus of contemporary American medical texts.* ICAME Journal 1998;22:73-80.

[18] Sinclair J. *Preliminary recommendations on text typology.* WWW page http://nicolet.ilc.pi.cnr.it/EAGLES/text-typ/texttyp.html, EAGLES (Expert Advisory Group on Language Engineering Standards), June 1996.

[19] Biber D. *Representativeness in corpus design*. In Linguistica Computazionale 1994;IX-X:377-408. Current Issues in Computational Linguistics: in honor of Don Walker.

[20] Dublin Core Metadata Inititative . *The Dublin Core element set version 1.1*. WWW page http://purl.org/dc/documents/rec-dces-19990702.htm, 1999.

[21] Ide N, Priest-Dorman G, and Véronis J. *Corpus encoding standard.* Document CES 1, MULTEXT/EAGLES, http://www.lpl.univ-aix.fr/projects/eagles/TR/, 1996.

Pierre Zweigenbaum

Medical Informatics Department, DSI/AP-HP

Department of Biomathematics, Paris 6 University

91, boulevard de l'Hôpital

75634 Paris Cedex 13 (France)

Tel : +33 (0)1 45 83 67 28

Fax : +33 (0)1 45 86 80 68

Email : pz@biomath.jussieu.fr

Web site : http://www.biomath.jussieu.fr/~pz/

ERRATUM, Newsletter Vol. 6 n° 3

In the previous issue of the ELRA newsletter, Patrick Paroubek's article, entitled "*An Expert Bird's Eye View on Evaluation in Speech and Language Engineering*", has been cut.

The last item should read as follows:

"[...] Note: The slides presented during the bullet course will be available at the URL: http://www.limsi.fr/TLP/CLASS/class_events.html."

We apologise to the author and to our readers for this mistake.

New Resources

ELRA-W0015 “Le Monde” text corpus

Year **2000** has been appended to the collection of text corpora from Le Monde newspaper.

Electronic archiving of "Le Monde" articles started on 1 January 1987. Some 200 articles are added every day, and as of October 1997 the database contains more than 500,000 articles, making it the biggest of its kind for all French daily newspapers.

The corpus is available in an ASCII text format. Each month consists of some 10 MB of data (circa 120 MB per year). Data ranging from 1987 until present date are available through ELRA (each buyer may purchase up to 5 years of data).

Prices for research use

	ELRA Members	Non Members
1 year	238.91 Euro	310.59 Euro
2 years	477.83 Euro	621.17 Euro
3 years	716.74 Euro	931.76 Euro
4 years	955. 65 Euro	1242.35 Euro
5 years	1194.56 Euro	1552.93 Euro

ELRA-S0115 American English SpeechDat-Car database

The American English SpeechDat-Car database comprises 314 American English speakers (150 males, 164 females) recorded over the mobile telephone network. The SpeechDat-Car database is owned by Siemens and has been collected by the Oregon Graduate Institute of Science and Technology, in a subcontract through ELDA. This database is partitioned into 94 CDs (or 13 DVDs). The speech databases made within the SpeechDat-Car project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat-Car format and content specifications.

The speech data files are in two formats. Four of the microphones were recorded on the computer in the trunck of the car. These are stored as 16 kHz, 16 bit and uncompressed. The fifth microphone was connected to the cell phone, and was recorded on a remote machine. The U.S. telephone network uses a digital encoding of 8bit, 8kHz, with Mu-law compression. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items:

- 2 voice activation keywords;
- 1 sequence of 10 isolated digits;
- 7 connected digits : 1 sheet number (5+ digits), 1 spontaneous telephone number, 3 read telephone numbers, 1 credit card number (14-16 digits), 1 PIN code (6 digits);
- 3 dates : 1 spontaneous date (e.g. birthday), 1 prompted date, 1 relative or general date expression;
- 2 word spotting phrases using an application word (embedded);
- 4 isolated digits;
- 7 spelled words : 1 spontaneous (own forename or surname), 1 spelling of directory city name, 4 real word/name, 1 artificial name for coverage;
- 1 money amount;
- 1 natural number;
- 7 directory assistance names : 1 spontaneous (own forename or surname), 1 city of birth / growing up (spontaneous), 2 most frequent cities, 2 most frequent company/agency, 1 "forename surname";
- 9 phonetically rich sentences;
- 2 time phrases : 1 time of day (spontaneous), 1 time phrase (word style);
- 4 phonetically rich words;

· 67 application words: 13 mobile phone application words, 22 IVR function keywords, 32 car products keywords

- 2 additional language dependent keywords;
- spontaneous sentences (for last 100 speakers).

The following age distribution has been obtained: 130 speakers are between 16 and 30, 101 speakers are between 31 and 45, 79 speakers are between 46 and 60, and 4 speakers are over 60.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA Members	Non Members
Price for research use	90,000 Euro	120,000 Euro
Price for commercial use	90,000 Euro	120,000 Euro

ELRA-W0029 Amaryllis Corpus for the Evaluation of Information Retrieval Systems

Launched at the end of 1995, the AMARYLLIS project aimed at evaluating information retrieval software for French text corpora in order to provide a methodology for the evaluation of other similar tools. AMARYLLIS was organised by the Institut de l'Information Scientifique et Technique (INIST) with the support of the Agence francophone pour l'enseignement supérieur et la recherche (AUPELF-UREF) and the French Ministère de l'Education Nationale, de la Recherche et de la Technologie (MERT).

More specifically, the objective was to create document corpora, questions and answers, in the framework of the Action de Recherche Concertée (ARC A1, renamed as Amaryllis- Access to text information in French), in order to get similar works to the United States project TREC.

For more information about the AMARYLLIS project, please visit the following web site: http://www.inist.fr/accueil/profran.htm

All corpora are structured as SGML files with isolatin character-encoding.

The available corpora were provided by:

- INIST (Institut de l'Information Scientifique et Technique)
- OFIL (Observatoire Français et International des Industries de la Langue)
- ELRA (European Language Resources Association)

Each provider provided three types of corpora : text documents, search topics and answers to these topics in the corresponding text corpora (with frames of reference for the answers).

1- Text documents in French

The text documents in French comprise:

- Articles (titles and texts) extracted from the newspaper “Le Monde”; each batch contains three months of documents, provided by OFIL (01-01-93/31-03-93, 01-04-93/30-06-93),
- Titles and summaries of scientific articles covering every domain from the Pascal bibliographical databases (from 1984 to 1995) and Francis (from 1992 to 1995), provided by INIST.

The tagging of the documents conforms to a simplified version of a DTD from the TEI, which includes the possibility to manage the logical structure.

2- Multilingual text documents

The multilingual text documents have been provided by ELRA, and comprise documents in 6 languages (French, English, Italian, Spanish, German and Portuguese), extracted from the parallel corpus MLCC which contains documents translated in official European languages (from 1992 to 1994). The corpus was divided in two sub-corpora: written questions (10 million words) and debates of the European Parliament (5 to 8 million words per language).

3- Search topics

The topics derive from questions asked by end users, and should contain every information which is necessary to understand the issue they deal with and to estimate the relevance. They comprise the following items:

- A domain, to determine the field of knowledge they belong to,
- A topic: which equals to a title defining the subject,
- A question: which matches the question the user may ask,
- Complementary information: which gives details on further documents that should be selected from the corpus,
- Concepts: which are a set of descriptors used to set the limits of the search.

The topics have been built by OFIL, by some documentalists working for Le Monde who used requests from journalists, and by engineers responsible for documentation at INIST (experts in their domain) who used requests from end users. These topics were to cover numerous application fields, and to get a large number of relevant results in each corpus. The topics have been tested on the corpora to control their relevance. The query may have had to be modified, or some further details may have been needed.

4- Frames of reference for the answers

Answers' files contain for each numbered topic the numbers of all relevant documents. Some frames of reference for the answers were established before the participants proceeded to the tests. The answers had been selected by the providers (OFIL and INIST) with the appropriate methodology and adequate tools (initial frames of reference): they proceeded to a pre-selection of documents as extended as possible, based not only on their titles and summaries but also on the keywords and classification codes used in the Pascal and Francis databases. These keywords and classification codes can not be accessed by the participants. The results (a set of documents) are sorted manually, so that the results match the best the query.

The initial frames of reference were checked manually by the providers (INIST and OFIL), using the answers given by the participants. These answers were collected when the tests were finished. This allowed us to review and correct the frames of reference for the answers in order to give some even more detailed information for their content.

The 4 CDs contain each a corpus for the two phases of the two campaigns which took place. TrecEval is also provided.

	ELRA Members	Non Members
Price for research use	45 Euro	45 Euro
Price for commercial use	100 Euro	100 Euro

ELRA-S0116 Italian SpeechDat(II) MDB-250

The Italian SpeechDat(II) MDB-250 database comprises 375 Italian speakers recorded over the Italian mobile telephone network. The database was produced by CSELT in Turin, Italy. The MDB-250 database is partitioned into 6 CDs in ISO 9660 format. The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information. Each speaker uttered the following items:

- 2 isolated digits;
- 1 sequence of 10 isolated digits;
- 4 connected digits: 1 prompt sheet number (5+ digits), 1 telephone number (9-11 digits), 1 credit card number (14-16 digits), 1 PIN code (6 digits);
- 3 dates: 1 spontaneous date (e.g. birthday), 1 prompted date (word style), 1 relative and general date expression;
- 1 word spotting phrase using an application word (embedded);
- 6 application words;
- 3 spelled words: 1 spontaneous name (own forename), 1 city name, 1 real / artificial word for coverage;
- 1 Lira currency money amount;
- 1 natural number;
- 7 directory assistance names: 1 spontaneous name (own forename), 1 city of birth / growing up (spontaneous), 2 most frequent cities (set of 25), 2 most frequent company / agency (set of 25), 1 'forename surname' (set of 150 'full' names);
- 2 questions including 'fuzzy' yes / no: 1 predominantly 'Yes' question, 1 predominantly 'No' question;
- 9 phonetically rich sentences;
- 2 time phrases: 1 time of day (spontaneous), 1 time phrase (word style);
- 4 phonetically rich words.

5 more items were added to the Italian corpus (4 spontaneous, 1 read).

The following age distribution has been obtained: 3 speaker are below 16 years old, 147 speakers are between 16 and 30, 149 speakers are between 31 and 45, 56 speakers are between 46 and 60, 48 speakers are over 60. A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA Members	Non Members
Price for research use	15,000 Euro	18,750 Euro
Price for commercial use	21,000 Euro	26,250 Euro

ELRA-S0117 Italian SpeechDat(II) FDB-3000

The Italian SpeechDat(II) FDB-3000 database comprises more than 3000 Italian speakers (1494 males, 1546 females) recorded over the Italian fixed telephone network. The database was produced by CSELT in Turin, Italy. The FDB-3000 database is partitioned into 6 CDs in ISO 9660 format, each CD contains the recordings of 550 speakers. The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information. Each speaker uttered the following items:

- 1 isolated digits;
- 1 sequence of 10 isolated digits;
- 4 connected digits: 1 prompt sheet number (5+ digits), 1 telephone number (9-11 digits), 1 credit card number (14-16 digits), 1 PIN code (6 digits);
- 3 dates: 1 spontaneous date (e.g. birthday), 1 prompted date (word style), 1 relative and general date expression;
- 1 word spotting phrase using an application word (embedded);
- 3 application words;
- 3 spelled words: 1 spontaneous name (own forename), 1 city name, 1 real / artificial word for coverage;
- 1 Lira currency money amount;
- 1 natural number;
- 5 directory assistance names: 1 spontaneous name (own forename), 1 city of birth / growing up (spontaneous), 1 most frequent cities, 1 most frequent company / agency, 1 'forename surname';
- 2 questions including 'fuzzy' yes / no: 1 predominantly 'Yes' question, 1 predominantly 'No' question;
- 9 phonetically rich sentences;
- 2 time phrases: 1 time of day (spontaneous), 1 time phrase (word style);
- 4 phonetically rich words.

4 more items were added to the Italian corpus.

The following age distribution has been obtained: 133 speaker are below 16 years old, 757 speakers are between 16 and 30, 862 speakers are between 31 and 45, 626 speakers are between 46 and 60, 482 speakers are over 60. A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA Members	Non Members
Price for research use	24,500 Euro	31,000 Euro
Price for commercial use	34,000 Euro	39,000 Euro

ELRA-S0118 Greek SpeechDat(II) FDB-5000

The Greek SpeechDat(II) FDB-5000 database comprises 5000 Greek speakers (2405 males, 2595 females) recorded over the Greek fixed telephone network. The database was produced by the Wire Communications Laboratory of the Department of Electrical Engineering at the University of Patras and Knowledge SA in Greece. The FDB-5000 database is partitioned into 25 CDs in ISO 9660 format. The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items:

- 2 isolated digits;
- 1 sequence of 10 isolated digits;
- 7 connected digits: 1 prompt sheet number (5+ digits), 1 telephone number (9-11 digits), 1 credit card number (14-16 digits), 1 PIN code (6 digits), 1 long number greater than 999999, 1 decimal number, 1 age (spontaneous);
- 3 dates: 1 spontaneous date (e.g. birthday), 1 prompted date (word style), 1 relative and general date expression;
- 1 word spotting phrase using an application word (embedded);
- 3 application words;
- 3 spelled words: 1 spontaneous name (own forename), 1 city name, 1 real / artificial word for coverage;
- 1 currency money amount;
- 1 natural number;
- 7 directory assistance names: 1 name (e.g. forename, spontaneous), 1 city of birth / growing up (spontaneous), set of 150 SDB full names , 1 most frequent cities, 1 most frequent company / agency, 1 city/region of call (spontaneous), 1 profession (spontaneous);
- 4 yes / no questions, 1 fuzzy yes/no question that could have either yes/no or something else as an answer;
- 9 phonetically rich sentences;
- 2 time phrases: 1 time of day (spontaneous), 1 time phrase (word style);
- 4 isolated words;
- 1 male/female (spontaneous);
- 1 telephone model (spontaneous);
- 1 environment of call (spontaneous);
- 5 words broken into syllables.

The following age distribution has been obtained: 512 speaker are below 16 years old, 2555 speakers are between 16 and 30, 1199 speakers are between 31 and 45, 653 speakers are between 46 and 60, 74 speakers are over 60, and 7 speakers age unknown.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA Members	Non Members
Price for research use	35,000 Euro	60,000 Euro
Price for commercial use	50,000 Euro	70,000 Euro

ELRA-S0114 Strange Corpus 10 - SC10 (‘Accents II’)

70 speakers (67 non-native, 3 native German speakers) - 1 dialogue, 1 re-telling of a German story - transliteration, orthography, canonical transcription. A collection of a variety of speech styles spoken by native and non-native German speakers, read texts, numbers, phonetically balanced sentences, story, free monolog, dialog.

	ELRA Members	Non Members
Price for research use	255.65 Euro	511.29 Euro
Price for commercial use	1255.22 Euro	1510.87 Euro

LREC 2002

Language Resources and Evaluation Conference

from 27th May 2002 to 2nd June 2002,

in Las Palmas, Canary Islands (Spain)

The conference topics will cover issues related to, on one hand, the design, the construction and the use of language resources *(mechanisms of language resources distribution and marketing, multimodal and multimedia resources, industrial use of language resources, or Internet-accessible metadata descriptions of language resources are but a few examples)*, and, on the other hand, to Human Language Technologies (HLT) evaluation *(e. g. benchmarking of systems and products; resources for benchmarking and evaluation, evaluation in written and spoken language processing, evaluation of multimodal systems, evaluation methodologies, protocols and measures, etc.)*.

ELRA Application form

Organisation .... Department ....

Name of Designated Representative..

Address .... Town .... Postcode ....

Country .... Telephone .... Fax ....

Email: ....Web...

College ( ) Spoken ( ) Written ( ) Terminology

Category:	( ) Non-profit-making organisations	750 EURO/year
	( ) European SME of less than 50 employees	1000 EURO/year
	( ) European profit making organisations of more than 50 employees	1500 EURO/year
	( ) Non European profit making organisations	5000 EURO/year

( ) I agree to the information above appearing in the ELRA Directory

Signature: Date:

For information, please contact:
ELRA, 55-57 rue Brillat Savarin - 75013 PARIS, FRANCE
Tel : +33 1 43 13 33 33 - Fax : +33 1 43 13 33 30 - Email: mapelli@elda.fr

Notes

1. An invoice for the membership fee will be sent upon receipt of the completed application form, and should be paid within 30 days.

2. Payment may be made by bank transfer or cheque in EURO, made out in favour of ELRA. Bank : BNP (Luxembourg) S.A,

24, Bd. Royal, L2953 Luxembourg
Account n°: 63-114418-57-6102-997.
Bank charges to be borne by the subscriber.

3. Membership covers the period from 1 January to 31 December of each year