The ELRA Newsletter

April - June 2001

Vol.6 n.2

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.

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,

As you may know, our General Assembly took place at the beginning of April. The Board members, some members of the Association, and all ELDA staff members attended the meeting. The CEO and the Board made a presentation of the ELRA and ELDA management and financial reports of the previous year, which were both unanimously approved. The activities, tasks, and missions ELRA/ELDA will be involved in this year were also presented.

As required by our status, the collection of the annual subscription fees was discussed. As for the amount of the fees, the General Assembly decided not to change them:

Non-profit making organisations	*750 EURO*
European small/medium-sized companies < 50 employees	*1000 EURO*
European profit making organisations >= 50 employees	*1500 EURO*
Non European profit making organisations	*5000 EURO*

The timing of the collection of the subscription fee has been set on 1st December 2001. You will receive the renewal invoice by then.

Among the activities planned for 2001 is the preparation of LREC 2002, which will constitute an important task for ELRA/ELDA. We have produced a first draft of the requirements that should be fulfilled to organise a successful conference. This document will be soon widely distributed to get feedback. The LREC 2002 conference will very likely take place in Las Palmas, Gran Canaria Islands, Spain, during the last week of May / first week of June 2002.

At the General Assembly, the Board announced its plan to hold a meeting to discuss the strategy of ELRA/ELDA for the two to five years to come. This meeting took place in May, and the major decisions are related to the involvement of ELRA/ELDA in the production/commissioning of Language Resources, the validation of LRs, the extension of the areas it covers to include multimodal/multimedia resources, and, last but not least, the start-up of a new activity regarding the evaluation (of technologies, systems, prototypes, services, etc.). A small group was also set up to work on international partnerships and cooperations.

In the field of speech resources, ELRA/ELDA is currently involved in the production of several databases: for consumer products in the framework of SPEECON, for Broadcast news in the framework of Network-DC, others are under preparation, in the framework of a new EU funded project called Orientel, and will cover the Arabic, Turkish, Hebrew and Greek (Cyprus) languages.

Other national or European projects and initiatives are to be launched later this year: AMARYLLIS, CLEF (Cross-Language European Forum), ISLE (International Standards for Language Engineering), HOPE 2001 (HLT Opportunity Promotion in Europe 2001), and ENABLER (European National Activities for Basic Language Engineering and Resources).

Now, considering the "internal" activity, the ELRA/ELDA web sites are being reworked and redesigned, and every service offered by ELRA/ELDA with their detailed description will be listed. Simultaneously, we are also reorganising the catalogue of LRs, thus reformatting the descriptions of the resources currently available so that the user can make a search with accurate criteria.

As for the content of this newsletter, the first part is a call for contributions from the partners of the ISLE NIMM (Natural Interactivity and MultiModal resources) project, who are collecting every kind of information on data resources and coding schemes for natural interactivity and multimodality.

A second article was written by Professor Shigeyoshi Kitazawa about the Japanese prosodic database which is currently being developped.

Finally, a third article deals with the European Terminology Information Server (ETIS). Suzanne Lervard, as a member of the European Association for Terminology (EAFT) Board wrote a paper to present this initiative in the field of terminology.

As usual, the final section of the newsletter is dedicated to the new resources available in our catalogue since the last issue in March, listed below:

ELRA-S0101: Spanish SpeechDat(II) FDB-1000
ELRA-S0102: Spanish SpeechDat(II) FDB-4000
ELRA-S0103: Swiss-French SpeechDat(M)
ELRA-S0104: Swiss-French SpeechDat(II) FDB-3000
ELRA-S0105: Swiss-German SpeechDat(II) FDB-2000
ELRA-S0106: Dutch SpeechDat(II) MDB-250
ELRA-S0107: Flemish SpeechDat(II) FDB-1000
ELRA-S0108: Belgian-French SpeechDat(II) FDB-1000
ELRA-S0109: Luxemburgish-French SpeechDat(II) FDB-500
ELRA-S0110: Luxemburgish-German SpeechDat(II) FDB-500

Please do not hesitate to contact us if you wish to make comments and suggestions or contribute a paper to the next issue, which is due to be published and distributed next September.

Antonio Zampolli, President

Khalid Choukri, CEO

Call for contribution: Multimodal and Natural Interactivity Corpora and Coding Schemes

Jean-Claude Martin

Natural interaction and multimodal resources concern the combination of several human communication modalities, such as speech, hand and body gesture, facial expression, gaze, etc. These resources take the forms of raw or annotated data (corpora), coding schemes for data annotation, or coding support tools, and are increasingly needed in many different areas of science, including human-computer interaction, systems development, communication research, linguistics, psychology, etc.

Partly to address those needs, the ISLE project on International Standards for Language Engineering (http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm) was launched in January 2000. ISLE aims to develop and promote Human Language Technologies (HLT) standards and guidelines for language resources, tools and products in key growth areas. The project has a Working Group in each of the following three areas: natural interaction and multimodality (NIMM), multilingual lexicons, and evaluation of machine translation systems. ISLE involves cross-Atlantic collaboration and is a continuation of EAGLES (Expert Advisory Group for Language Engineering Standards, http://www.ilc.pi.cnr.it/EAGLES96/browse.html).

The ISLE NIMM Working Group (http://isle.nis.sdu.dk/) is in the process of surveying and developing standards and guidelines for natural interaction and multimodal data resources, annotation schemes and coding tools, as well as meta-data descriptions for large multimodal databases. ISLE NIMM consists of nine partners from seven European countries.

The ISLE NIMM website already includes a survey of coding tools for natural interactivity and multimodality (under Reports at http://isle.nis.sdu.dk/). We are presently collecting information for world-wide dissemination at this website on data resources and coding schemes for natural interactivity and multimodality. You are cordially invited to provide us with a description of a data resource which you have constructed, a coding scheme which you are using for annotating a multimodal or natural interactivity corpus, or an annotation tool you have developed for coding multimodal or natural interactivity corpora. To do so, please use the ISLE NIMM forms available on the Reports 8.1, 9.1 and 11.1 pages, respectively, under Results at http://isle.nis.sdu.dk/. For any other questions or comments please contact the authors by email.

Web site: http://isle.nis.sdu.dk/

Niels Ole Bernsen
Email: nob@nis.sdu.dk

Laila Dybkjaer
Email: laila@nis.sdu.dk

Malene Wegener
Email: mwk@nis.sdu.dk

Natural Interactive Systems Laboratory (NISLab)
University of Southern Denmark-Odense-Science Park 10
DK-5230 Odense M
(Denmark)

Jean-Claude Martin
Email: martin@limsi.fr

Laboratoire d'Informatique et de Mécanique pour les Sciences de l'Ingénieur (LIMSI)
Centre National de Recherche Scientifique (CNRS)
BP 133
91403 Orsay cedex
(France)

and

Laboratoire d'Informatique et de Communication (LINC)
Institut universitaire de Technologie de Montreuil (Paris 8)
140, rue de la Nouvelle France
93100 Montreuil
(France)

Preparation of a Japanese Prosodic Database

Shigeyoshi Kitazawa

A new speech database project was initiated in Japan in October 2000, with a planned term of four years. Focusing on "The Realization of Advanced Spoken Language Information Processing from Prosodic Features", the project is headed by Professor Keikichi Hirose, Department of Frontier Informatics, School of Frontier Sciences, The University of Tokyo.

This new project is divided into several subgroups with research areas including theoretical, phonological, pathological, interactive, discourse, as well as speech recognition and synthesis. A large number of university researchers from computer science departments, linguistic, psychological, and medical departments, as well as some company-based researchers are cooperating to further this research.

The main theme of the project is "The Realization of advanced spoken language information processing from prosodic features"

In most research on speech technology, prosodic aspects of speech have been relatively little considered. However, as prosodic features play an important role in transmitting various kinds of information, their systematic and intensive investigation is necessary. From this point of view, the basics of prosody will be studied, including prosodic modeling, prosodic variation, and prosodic corpora. Improvements in speech synthesis and recognition will also be realized through research on prosodic aspects. Furthermore, spoken dialogue systems and medical welfare will also be studied with respect to prosody. This project will not only promote research on prosody but also make a drastic advancement possible in spoken language information processing technology.

Teams and subjects

Administration group

In order to realize further advancements in spoken language technology, a unified scientific study with a clear and well-formulated purpose should be conducted on prosody. From this point of view, our project will be organized with the aim of promoting research work on all aspects of prosody, from basics to applications. As the administration group, we handle various affairs for a smooth realization of the project, including regulation between research groups, organization of academic meetings, publication of research outputs, etc.

Analysis of prosody, its formalization and modeling

(Hiroya Fujisaki, Science University of Tokyo)

Prosodic relations, such as basic frequency patterns, utterance speed, tempo, rhythm, etc., are not yet well-enough understood in quantitative terms. The information expressed by prosody will therefore be classified under the following three categories:

- 1 Linguistic information (including vocabulary, syntactic structure, and discourse structure),
- 2 Paralinguistic information (including speaker's intention and attitude),
- 3 Non-linguistic information (including speaker's individuality, feelings, and individual differences).

Our research aims at modeling the processes which govern the prosodic-feature variations, in a form which can be used for voice information processing, after having first analyzed the prosodic-features, especially the fundamental frequency pattern and the time structure, and establishing clearly their relationships.

Variability of prosody and its quantitative expression

(Masuzo YANAGIDA, Doshisha University)

To develop feature parameters for quantitative description of prosody in real natural speech, trying to collect speech in real situations including emotional speech. Also to investigate individuality in prosody, situation effects, regional effects, perception of prosody, linguistic study of prosody.

Design of prosodic corpora and automation of the developing process

(Shigeyoshi Kitazawa, Shizuoka University)

This group will provide several different types of prosodic databases to meet requirements of research from different fields: text reading, semi-natural utterances, imitation dialog speech with an expression picture, and prosodic corpus which gave prosodic information to the existing speech corpus.

Prosody control for high-quality speech synthesis

(Keikichi Hirose, The University of Tokyo)

The aim of the group is to establish an advanced synthesis technology of prosodic features, enabling us to generate synthetic speech highly human-like in various utterance styles. The realization of a spoken dialogue system, which replies with synthetic speech which will be highly acceptable for users, is also aimed at. Both of heuristic and statistical frameworks are introduced for the research works.

Use of prosodic information in speech recognition and understanding

(Kazuhiko Ozeki, The University of Electro-Communication)

This group is concerned mainly with linguistic information contained in prosodic features of Japanese utterances, and exploiting it for various problems in speech recognition and understanding. Examples of the challenges are: enhancement of accuracy and efficiency in speech recognition by using prosody, use of prosody for Japanese dependency analysis, construction of a speech recognition system with combined use of phonemic information and accent information, and detection of focus and prominence for use in summarization of spoken materials.

Discourse system with enhanced prosody control

(Tetsunori Kobayashi, Waseda University)

This group investigates the relation of the various prosodic-features and the dialog phenomena. A dialog system in which a natural dialog progression is possible will be built by use of prosody. For this purpose, we will develop the research in the following stages: construction of the model of prosody control of the dialog based on analysis of the relation of the natural flow and natural prosody of a dialog including a turn-taking, development of the prosody use policy in the dialog recognition from relation between an utterance act and a correction, modeling of a speaker shift from the viewpoint of prosody in two or more speakers conversation, and then we will build and improve a system.

Applications of prosody processing technology to the field of medicine and welfare

(Hideki Kasuya, Utsunomiya University)

The project concentrates on the prosody disorders that are typically accompanied by the motor speech disorders and voice disorders. In parallel with basic understanding of the physiological and physical mechanisms of prosody control, acoustic properties of speech utterances of the

patients are examined. The project's final goal is to explore a useful means for the examination, evaluation and enhancement for the prosody disorders as well as to develop aids for the patients.

Planned Prosodic Databases:

Reading and semi-natural utterances of Japanese MULTEXT

(Shigeyoshi Kitazawa, Shizuoka University)

Phoneme labels, syllable/mora labels, and prosodic structure labels (prosodic phrase: prosodic clause: prosodic sentence: prosodic word:) will be aligned to the read speech and to the semi-natural speech of the MULTEXT Japanese corpus, read by three men and women, and collected during the fiscal year 2000 (with simultaneous recording of speech and EGG signals). In addition, stylized intonation curves will be extracted through the manual correction by analysis and synthesis in conjunction with perceptual evaluation. Moreover, the directed paralinguistic information will be assigned to these imitated natural utterances.

Imitated dialog speech using pictures

(Akira Ichikawa, Chiba University)

Voice data recording using expressive pictures is improved and a procedure for the recording of speech data using expressive pictures is going to developed. A form of dialog corpus with an expression picture and creation soft specification is examined. The recording of a natural dialog speech and corpus will be used for developing methods for automatic phoneme segmentation and evaluation of the data.

Attachment of prosodic information to the existing speech corpus

(Shuichi Itahashi, University of Tsukuba)

This is based on investigation of prosodic corpus in and outside the country, the prosody declaring method is compared and examined, and the optimum declaring method is defined. Based on it, prosodic analysis of continuous speech data (about 12500 sentences from transcribed free conversations) of Japanese Acoustics Society speech corpus used widely is carried out.

Preliminary Corpus:

As a preliminary stage of the project, a trial Prosody Corpus based on the framework of MULTEXT is being produced. With the cooperation of the Tokyu Construction Technical Research Institute, we have finished mastering recordings taken from three men and three women including professional narrators and actors and actresses.

The prosodic features of the data recorded for this project differ from those of conventional voice databases in the following ways:

1- It was recorded using precision equipment for sound measurement in a perfect non-reverberant room. Dark noise, reverberation characteristics and all remaining noises of the measurement system etc. were measured, and conditions were adjusted precisely to be exactly the same at each recording session.

2- All recordings include signal from EGG electrodes, recorded simultaneously with the microphone signal.

3- The text was prompted and uttered exactly as written. Two kinds of utterances, reading and performance, were recorded. For all speakers and for both styles of reading, speaking style was scripted and controlled so that the same accent, prominence, breathing, etc., were obtained. Also in performance versions, accents and prominences were constrained to be constant, and breaths and pauses were also controlled.

4- The text consists of 40 short paragraphs which were obtained by functionally equivalent translation from the original MULTEXT sources. They consist of various topics, situations, and styles, including, for example, telephone orders, telephone complaints, urgent reports, telephone reference, a presentation, a general report, traffic information, an apology, a boast, a letter, occasional thoughts, a discourse, a lecture, sections of a novel, a monologue, etc. Although some sentences were self-contained, each passage presupposed that pauses or silences between sentences are to be included as data in the recordings when a whole paragraph is spoken without mistake in one session. When mistakes were made in the performance, the whole passage was restarted from the beginning.

5- The imitation of natural utterance by performance can not be obtained on-demand by simple instructions such as saying "please speak freely", so each situation is described beforehand, and the speaker has to perform in the same image, with repeated attempts at naturalness until the supervisor of the recording session was able to judge that the image has been adequately expressed. In addition, the performances were directed and controlled so that the accents, prominences, breathings, pauses, timing and rhythm are constrained to be the same as that produced at the time of reading. Since these details differ from speaker to speaker, and are noted at the time of recording, there may be individual differences between speakers.

6- The Japanese translations were produced from the MULTEXT originals after taking into account the linguistic and cultural variations of the 5 original languages. A unified text was produced, and certain modifications were made with respect to the Japanese situation. The tone of some sentences is a bit long for an oral utterance, and there is a hardness arising from the translation. The texts also have the features (see note 1) of written language and therefore of expressing the more formal features of written text rather than spontaneous speech.

Tagging is necessary, but there is not yet general a agreement about which method should be used. J-ToBI labeling has been proposed, but not enough researchers have had experience with this system for it to be the natural choice. We believe that it is necessary to mark accentual phrases, accent kernels, prosodic clause boundaries, intonation type (flat, ascending, descending), prominence, focus, lengthening, and devocalization. In addition, morphological, lexical, syntactic, and discourse labeling will be required, as well as paralinguistic labeling. Although some researchers have argued that these labels represent merely the interpretation of the labeler and are not necessary reliable, we believe that information should be provided at the level of the morpheme, the word, the phoneme, the clause (including clause dependency), the sentence (including speaker turns), and the emotional state if possible. It is also useful to have details about the dialog partners' relationship, language variant, and dialect when possible.

Notes 1:

Although some participants questioned the recording of such texts and considered rewriting them to a style more appropriate to spoken-language, e.g., omitting the subject where possible to be consistent with Japanese speech, the consequent results indicated further problems - e.g., there are numerous modification patterns required for the different speaking styles of men and women in Japanese, and in the case of student language, where status and age differences should be considered. Such cultural differences appear endless. Abbreviations are performed naturally in Japanese, and the exact meaning of a text is rarely transmitted. Utterances become

fragmentary and tendency is for reduction to a simple prosody pattern. The text of our MULTEXT corpus is therefore not representative of spoken language, but is instead the written language equivalents. The MULTEXT documents say that when having been read, natural intonation (that is, calm usual voice) was requested of the informants. The sound actually recorded is heard as fluently and with natural speed. There is no particular feeling of tension, and the texts were spoken in a neutral attitude, without any expression of emotional feeling.
We consider the readings to be appropriate to the intention of having translated and recorded the texts of 5 languages, and have produced a translation which expresses the same paralinguistic information. i.e., not machine translation but situation-dependent equivalence. The contents may therefore be changed suitably so that the actual conditions of each country may be matched in terms of emotion, and feeling. The purpose of MULTEXT was not described clearly, and the project has already ended so no connection can be made with the original persons concerned. Although the distributed MULTEXT recorded sound had aimed at neutral fluent sound which read aloud, it is surmised that recording of paralinguistic information is left as future work.
Taking the above considerations into account, our text does not aim at a spoken-language tone, but is a written tone in the neutral style (between broken colloquial speaking style and formal stiff-mannered styles), it considered as appropriate language to express the meaning of the original text of the EUROM1 as closely as possible. When detailed supplementary information is added in each passage in the five languages, the sum-set was considered, details are written and added so that the passage might become explanation as detailed as possible.
The completed Japanese text contains complicated word dependencies. Therefore, although cautions in prosody were required for each utterance and many were hard for narrators to read naturally, as a result abundant prosodic data was collected. Thus, the uttered sound was the favor of the performed person's expression power, unnaturalness was not felt but there was no sense of incongruity, although the text was Japanese of a bit hard expression.

Dr. Shigeyoshi Kitazawa
Professor
Department of Computer Science,
Faculty of Information
Shizuoka University
3-5-1, Johoku, Hamamatsu-shi
Shizuoka-ken 482-8011 (Japan)

Tel.: +81-53-478-1471
Fax : +81-53-478-1499

Email: kitazawa@cs.inf.shizuoka.ac.jp

ETIS2 - The European Terminology Information Server: www.etisnet.net

Suzanne Lervard

The **E**uropean **T**erminology **I**nformation **S**erver (ETIS) is a server containing information about the field of terminology (calendar, biographical and bibliographical information, terminological theses, etc.), but it does not itself propose terminology. The data in ETIS are provided by the European Association for Terminology (EAFT) and the Terminology Documentation Center Network (TDCNet). ETIS is intended to be an open information tool dealing with all types of information concerning terminology activity in Europe. To this end, ETIS provides:
1. A harmonised interface for multi-site consultation of heterogeneous databases (the different bibliographic and factual databases of the terminology documentation centres (TDC's) in Europe;
2. Access to bibliographic data as well as activity-, institution- and expert-specific data;
3. Information in the field of terminology by offering a selection according to certain types of data.
Some of this information has been recorded and supplied in a comprehensive fashion, while other information consists of a representative selection drawn from large distributed data collections. ETIS is multilingual (15 languages represented) and contains links to other terminology servers - mainly those from the 16 partners in the TDCNet project:

-INFOTERM (international information centre for terminology),
-ASS.I.TERM (associazone italiana per la terminologia),
-CINDOC (centro de informacion y documentacion cientifica),
-CTB (centre de terminologie de Bruxelles de l'Institut Libre Marie Haps,
-CTN (centre de terminologie et de néologie),
-DANTERM (DANTERMcentret),
-DEUTERM (Deutsches Informations- und Dokumentationszentrum),
-ELOT (Hellenic organization for standardization),
IM (Islensk malstöd, Icelandic Language Institute),
-NTU (Nederlandse taalunie),
-TNC (Swedisch centre for technical terminology),
-TSK (tekniikan sanastokeskus ry),
-UL/DTIL (union latine),
-UZEI (basque centre for terminology and lexicography).

The languages are as follows: Catalan (ca), German (de), Danish (da), Greek (gr), English (en), Spanish (es), Finnish (fi), French (fr), Icelandic (is) Italian (it), Dutch (nl), Norwegian (no), Portuguese (pt), Swedish (sv).
The translations have been provided by the different documentation centres in the TDCNetproject.
From ETIS, the user can connect to the distributed databases maintained by these terminology documentation centres (TDCs) to obtain more detailed or additional information, depending on the type of data required.
ETIS is accessible from the TDCnet site: *www.tdcnet.net.*
The main aims of the TDCNet System are to pool resources of national centres and to make them as widely available as possible.
The documentation exchanged in this project is secondary or tertiary (i.e., bibliographies of dictionaries, theoretical works, standards, and bibliographies), as well as factual information, including notably announcements of teaching and training

opportunities, terminology projects, and a "who's who" in terminology. ETIS grew out of a recommendation from the POINTER project and was then drawn into the EAFT in the context of a Special Interest Group (SIG). The EAFT was founded on 3rd October 1996, in Kolding, Denmark. The EAFT is a non-profit organisation which aims to bring together any persons or organisations with an interest in terminology. The web site of the EAFT is: www.eaft-aet.net.

The EAFT intends to bring together all the individuals and institutions in Europe who are active in or have an interest in the discipline of terminology. In this context, "Europe" is interpreted broadly and is not limited to the member states of the European Union. The EAFT plans to develop co-operation agreements which will allow other institutions, networks or associations (from Europe and elsewhere) to participate in the EAFT. Such agreements have already been established with the International Information Centre for Terminology (Infoterm), Realiter (Pan-Latin terminology network), and the European Language Resources Association (ELRA).

ETIS is a window into the EAFT and into the different national and regional associations belonging to the EAFT (e.g. NL-TERM, DANTERM). In addition, ETIS also functions as a tool for the TDCNet, the external gateway of the extranet created by the TDCNet: all information intended for dissemination (i.e., all information excluding that which is confidential, under copyright, or tied to commercial interests) from the various terminological documentation centres that make up the TDCNet is disseminated via ETIS.

The development from ETIS 1 to ETIS 2 was a result of the recommendations from workpackage 2.2 of the TDCNet Consortium Project, which consisted of 6 workpackages such as the evaluation of the exiting documentation of the TDC's, some recommandations, the creation of a central database and the promotion of the activities, and specified certain rules concerning the presentation of data and has issued directives with regard to the format which should be adapted. As a result of this, ETIS is no longer a server that presents textual information (as was the case for the first version of ETIS), but one which offers access to databases (organized by country, by domain, by language, etc.), and which allows direct (rather than page by page) consultation.

ETIS also acts as the TDCNet user interface accessible at www.etisnet.net. The TDCNet website www.tdcnet.net has been designed in such a way that it also served as a promotion mechanism.

Browsing and accessing multilingual information

ETIS is designed as a multilingual interface in which the user chooses his or her language of consultation. Thus, each screen has to be able to deliver the necessary information in the chosen consultation language. In the following example, a Spanish source is displayed using the English and French interfaces.

The labels Title, Languages of the collection and Subject, classification are translated but not the content of the field Title.

English version:
Title: Terminologia informatica.
Languages of the collection: Spanish, German, English.
Subject, classification: Software, informatics.

French version:
Titre: Terminologia informatica.
Langues de la collection: espagnol, anglais, français.
Domaine, classification : Logiciel, informatique.

The main menu of ETIS offers the possibility to conduct searches from two main types of documents: Bibliographic data and Factual data.

The bibliographic data contain the following types:
- Literature in Serial Form
- Literature in Monographic Form
- Term Collections

The factual data type comprises six subtypes which can be examined separately:
- Corporate entities
- Projects
- Teaching and training opportunities
- Persons, experts
- Terminology Management Software
- Events

Accessories to the ETIS interface

In the upper part of the screen there are links to the TDCNet and EAFT and Partners in order to learn more about the "owners" of ETIS. On the left side of the screen of the main menu you can access further information: **About ETIS**, the **Main search** menu, the **Agenda**, **Glossary**, **Cart**, **Contact Us**. The glossary, for instance, contains all the terms and definitions that allow users to have a better understanding of the concepts used in ETIS.

OPTIMISATION OF ETIS 2

The second version of ETIS needs to be optimised in the future in the following ways:

· ETIS's integration in the metasite
· Study of new features
· Identification of alternative sources and means of import
· Integration, testing and evaluation of the data

The implementation of such features will allow ETIS to develop into the main reference point for European terminology information and to provide a very wide spectrum of information and links.

For more information on the goals and activities of the EAFT, please contact:
Rute Costa & Daniel Prado
Union latine, DTIL
131, rue du Bac
75340 Paris Cedex 07 (France)

Tel.: +33 1 45 49 60 60
Fax: +33 1 45 44 45 97

Web site of the AET: www.eaft-aet.net
Email: eaft_aet@unilat.org

Suzanne Lervard,
Member of the EAFT Board - European Association for Terminology

Email: lervad@easynet.fr

New Resources

ELRA-S0101 Spanish SpeechDat(II) FDB-1000

The Castillian Spanish SpeechDat(II) FDB-1000 comprises 1000 Castillian Spanish speakers (481 maleand 519 femal) recorded over the Spanish fixed telephone network. The SpeechDat database has been collected and annotated by the Department of Signal Theory and Communications of the Universitat Politècnica de Catalunya (UPC) (Spain). The FDB-1000 database is partitioned into 4 CDs, each of which comprises 250 speakers sessions. 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:
- 3 application words
- 1 sequence of 10 isolated digits
- 4 connected digits: 1 sheet number (6 digits), 1 telephone number (9-11 digits), 1 credit card number (14-16 digits), 1 PIN code (6 digits) (set of 150)
- 3 dates: 1 spontaneous date, e.g. birthday, 1 prompted date, word style, 1 relative and general date exp
- 1 word spotting phrase using an application word (embedded)
- 1 isolated digit
- 3 spelled word (letter sequences): 1 spelling of surname (same as O1), 1 spelling of direct. city name (O3), 1 real/artificial for coverage
- 1 currency money amount
- 1 natural number
- 5 names extracted from the telephone directory: 1 surname (set of 500), 1 city of birth / growing up (spont), 1 most frequent cities (set of 500), 1 most frequent company/agency (set of 500),1 forename surname (set of 150).
- 2 yes/no questions: 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

The following age distribution has been obtained: 19 speakers are below 16 years old, 555 speakers are between 16 and 30, 198 speakers are between 31 and 45, 198 speakers are between 46 and 60, and 30 speakers are over 60.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA Members	Non Members
Price for research use	9,000 Euro	22,000 Euro
Price for commercial use	18,000 Euro	25,000 Euro

ELRA-S0102 Spanish SpeechDat(II) FDB-4000

The Castillian Spanish SpeechDat(II) FDB-4000 consists of 4000 Castillian Spanish speakers (2061 males, 1939 females) over the Spanish fixed network, stored on 14 CD-ROMs in the final SpeechDat(II) database exchange format. Collection was performed at the Department of Signal Theory and Communications of the Universitat Politècnica de Catalunya (UPC) (Spain). 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.

The following items were recorded:
- 3 application words
- 1 sequence of 10 isolated digits
- 4 connected digits: 1 sheet number (6 digits), 1 telephone number (9-11 digits), 1 credit card number (14-16 digits), 1 PIN code (6 digits) (set of 150)
- 3 dates: 1 spontaneous date, e.g. birthday, 1 prompted date, word style, 1 relative and general date exp
- 1 word spotting phrase using an application word (embedded)
- 1 isolated digit
- 3 spelled word (letter sequences): 1 spelling of surname (same as O1), 1 spelling of direct. city name (O3), 1 real/artificial for coverage
- 1 currency money amount
- 1 natural number
- 5 names extracted from the telephone directory: 1 surname (set of 500), 1 city of birth / growing up (spont), 1 most frequent cities (set of 500), 1 most frequent company/agency (set of 500), 1 forename surname (set of 150).
- 2 yes/no questions: 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

The following age distribution has been obtained: 42 speakers are under 16 years old, 2234 are between 16 and 30, 844 are between 31 and 45, 764 are between 46 and 60, and 116 are over 60.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA Members	Non Members
Price for research use	28,000 Euro	48,000 Euro
Price for commercial use	40,000 Euro	56,000 Euro

ELRA-S0103 Swiss-French SpeechDat(M)

The Swiss-French SpeechDat(M) project comprises 1000 recorded Swiss-French speakers (575 female and 425 male speakers). The corpus contains phonetically rich sentences & application oriented utterances such as keywords, digits, etc..

The recording site was located at Bern, and collections were performed by the Swiss PTT on the SwissNet. The recordings were at first stored on CD-ROM using 8 bit A-law, and then sent to IDIAP for further processing. Speech samples are stored as sequences of 8-bit 8 kHz A-law speech samples (before compression). 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.

The following items were recorded:

- 1 sequence of 6 single digits including the hash (#) and the star (*) symbols;
- 1 sheet id number (5 connected digits / 1 natural number);
- 1 telephone number (spontaneous);
- 1 16-digit credit card number;
- 2 natural numbers (1 + sheet id);
- 2 money amounts;
- 1 quantity;
- 3 spelled words;
- 1 time phrase (prompted, word style);
- 1 date (spontaneous);
- 1 date (prompted);
- 1 yes/no question;
- 1 city name (prompted);
- 1 city name (spontaneous);
- 5 function words;
- 1 name (spelling table);
- 1 mother tongue (spontaneous);
- 1 education level (out of 3 choices);
- 1 telephone type (out of 6 choices);
- 10 sentences (read);
- 1 query to telephone directory (given the name and the city of subject);
- 1 free comment on the session.

	ELRA Members	Non Members
Price for research use	9,000 Euro	16,000 Euro
Price for commercial use	16,000 Euro	20,000Euro

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

ELRA-S0104 Swiss-French SpeechDat(II) FDB-3000

The Swiss-French SpeechDat(II) FDB-3000 comprises 3000 Swiss-French speakers (1500 males, 1500 females) recorded over the Swiss fixed telephone network. The recordings were performed by the Swiss PTT in Bern. This database is partitioned into 6 CDs, each of which comprises 500 speakers sessions. 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.

The following items were recorded:

- 5 application words;
- 1 sequence of 6 isolated digits including the hash (#) and the star (*);
- 3 connected digits: 1 sheet number, 1 telephone, 1 credit card number (16 digits);
- 2 dates: 1 spontaneous date, e.g. birthday, 1 prompted date, word style;
- 3 spelled words from a list of name and titles;
- 2 currency money amounts;
- 2 numbers: 1 natural number, 1 quantity number (prompted);
- 1 place (province of longest residence);
- 7 optional item: 1 name (spelling table), 1 city name, 1 mother tongue of speaker (spontaneous), 1 education level of speaker (out of 3 choices), 1 type of telephone used, 1 query to telephone directory;
- 1 free comment on session;
- 1 yes/no question;
- 10 phonetically rich sentences;
- 1 time phrase (word style).

The following age distribution has been obtained: 69 speakers are below 16 years old, 1006 speakers are between 16 and 30, 944 speakers are between 31 and 45, 629 speakers are between 46 and 60, 311 speakers are over 60, and 41 speakers whose age is unknown.

	ELRA Members	Non Members
Price for research use	25,000 Euro	34,000 Euro
Price for commercial use	32,000 Euro	38,000 Euro

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

ELRA S0105 - Swiss-German SpeechDat(II) FDB-2000

The Swiss-German SpeechDat(II) FDB-2000 comprises 2000 Swiss-German speakers (992 males, 1008 females) recorded over the Swiss fixed telephone network. The recording were performed by the Swiss PTT in Bern. This database is partitioned into 6 CDs. 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.

The following items were recorded:

- 3 application words;
- 1 sequence of 9 isolated digits including the hash (#) and the star (*);
- 1 sequence of isolated digits - only digits which are not representing in B1;
- 3 connected digits: 1 area code, 1 spontaneous phone number, 1 credit card number (16 or 15 digits);
- 3 dates: 1 spontaneous date, e.g. birthday, 2 prompted dates;
- 3 word spotting phrases using an application word (embedded);
- 1 isolated digit;
- 4 spelled words from a list of proper names and cities;
- 1 currency money amount;
- 1 natural number;
- 1 place of education (spontaneous);
- 1 type of telephone used (spontaneous);
- 1 query to telephone directory (spontaneous);
- 2 yes/no questions: one about smoker/non smoker and another about sex gender.;
- 9 phonetically rich sentences;
- 2 time phrases: 1 time of day (spontaneous), 1 time phrase (word style);
- 4 phonetically rich words.

The following age distribution has been obtained: 33 speakers are below 16 years old, 565 speakers are between 16 and 30, 623 speakers are between 31 and 45, 442 speakers are between 46 and 60, and 337 speakers are over 60.

	ELRA Members	Non Members
Price for research use	20,000 Euro	26,000 Euro
Price for commercial use	32,000 Euro	38,000 Euro

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

ELRA-S0106 Dutch Speechdat(II) MDB-250

The Dutch SpeechDat(II) MDB-250 comprises 250 Dutch speakers (125 males, 125 females) recorded over the Dutch mobile telephone network. The recordings were made at SPEX, the Netherlands, and the recording application was developed and run with Show 'N Tel. This database is partitioned into 5 CDs The speech databases made within the SpeechDat(II) project were validated by SPEX 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.

The following items were recorded:

- 8 application words (2 optional);
- 2 isolated digits;
- 1 sequence of 10 isolated digits;
- 3 connected digits: 1 telephone number (1-10 digits), 1 credit card number (1-16 digits), 1 digit PIN code (6 digits);
- 3 dates: 1 spontaneous date, 1 date, 1 relative date expression;
- 1 embedded application word
- 3 spelled words: 1 forename (spontaneous), 1 city name, 1 word;
- 1 currency money amount;
- 1 natural number;
- 6 directory assistance names: 1 forename (spontaneous), 1 city of birth, 1 most frequent city, 1 city name, 1 company name, 1 forename surname;
- 2 yes/no questions: 1 predominantly "yes" question, 1 predominantly "no" question;
- 9 phonetically rich sentences;
- 2 time phrases: 1 time of day (spontaneous), 1 time phrase;
- 4 phonetically rich words.

The following age distribution has been obtained: 5 speakers are under 16, 90 are between 16 and 30, 89 between 31 and 45, 56 between 46 and 60, and 10 are over 60.

	ELRA Members	Non Members
Price for research use	20,000 Euro	28,000 Euro
Price for commercial use	25,000 Euro	35,000 Euro

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

ELRA S0107 - Flemish SpeechDat(II) FDB-1000

The Flemish SpeechDat(II) FDB-1000 comprises 1023 Flemish speakers (461 Males, 562 Females) recorded over the Belgian fixed telephone network. Each phrase or word was repeated about 5 times.

The recordings were performed by Lernout & Hauspie, in Flanders (Belgium). This database is partitioned into 4 CDs, each of which comprises 250 speakers' sessions. 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.

The following items were recorded:

- 7 application words;
- 4 isolated digits;
- 1 sequence of 10 isolated digits;
- 5 connected digits: 1 area code, 1 spontaneous phone number, 1 credit card number (15-16 digits), 1 sheet number ; etc ;
- 3 dates: 1 spontaneous date, e.g. birthday, 1 prompted date, 1 general and relative date expression ;
- 1 embedded application word;
- 4 spelled words;
- 1 currency money amount;
- 1 natural number;
- 6 names extracted from the telephone directory: 1 surname, 1 city of birth, 1 most frequent cities, 1 most frequent companyu/agency, 1 forname-surname, etc;
- 4 yes/no questions: 1 predominantly yes question, 1 predominantly no question;
- 10 phonetically rich sentences;
- 2 time phrases: 1 time of day (spontaneous), 1 time phrase;
- 5 phonetically rich words.

The following age distribution has been obtained: 22 speakers are below 16 years old, 387 speakers are between 16 and 30, 306 speakers are between 31 and 45, 240 speakers are between 46 and 60 and 68 are over 60.

	ELRA Members	Non Members
Price for research use	14,400 Euro	25,200 Euro
Price for commercial use	18,000 Euro	25,200 Euro

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

ELRA S0108 - Belgian-french SpeechDat(II) FDB-1000

The Belgian-French SpeechDat(II) FDB-1000 comprises 1011 Belgian-French speakers (493 Males, 518 Females) recorded over the Belgian fixed telephone network. Each phrase or word was repeated about 2 times. The recordings were performed by Lernout & Hauspie, in Walonia (Belgium). This database is partitioned into 4 CDs, each of which comprises 250 speakers' sessions. 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.

The following items were recorded:

- 7 application words;
- 4 isolated digits;
- 1 sequence of 10 isolated digits;
- 5 connected digits: 1 area code, 1 spontaneous phone number, 1 credit card number (15-16 digits), 1 sheet number ; etc ;
- 3 dates: 1 spontaneous date, e.g. birthday, 1 prompted date, 1 general and relative date expression ;
- 1 embedded application word
- 4 spelled words;
- 1 currency money amount;
- 1 natural number;
- 6 names extracted from the telephone directory: 1 surname, 1 city of birth, 1 most frequent cities, 1 most frequent company/agency, 1 forename surname, etc ;
- 4 yes/no questions: 1 predominantly yes question, 1 predominantly no question;
- 10 phonetically rich sentences;
- 2 time phrases: 1 time of day (spontaneous), 1 time phrase ;
- 6 phonetically rich words.

The following age distribution has been obtained: 13 speakers are below 16 years old, 257 speakers are between 16 and 30, 425 speakers are between 31 and 45, 229 speakers are between 46 and 60 and 87 are over 60.

	ELRA Members	Non Members
Price for research use	14,400 Euro	25,200 Euro
Price for commercial use	18,000 Euro	25,200 Euro

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

ELRA-S0109 Luxemburgish-French SpeechDat(II) FDB-500

The Luxembourgish-French SpeechDat(II) FDB-500 comprises 614 Luxembourgish-French speakers (246 Males, 368 Females) recorded over the Luxembourgish fixed telephone network. Each phrase or word was repeated about 3 times. The recording were performed by Lernout & Hauspie, in Luxembourg. This database is partitioned into 3 CDs, each of which comprises 200 speakers' sessions. 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.
The following items were recorded:
- 7 application words;
- 1 sequence of 10 isolated digits;
- 4 isolated digits;
- 5 connected digits: 1 area code, 1 spontaneous phone number, 1 credit card number (15-16 digits), 1 sheet number ; etc;
- 3 dates: 1 spontaneous date, e.g. birthday, 1 prompted date, 1 general and relative date expression;
- 1 embedded application word;
- 3 spelled words;
- 1 currency money amount;
- 1 natural number;
- 6 names extracted from the telephone directory: 1 surname, 1 city of birth, 1 most frequent cities, 1 most frequent company/agency, 1 forename surname, etc;
- 4 yes/no questions;
- 10 phonetically rich sentences;
- 2 time phrases: 1 time of day (spontaneous), 1 time phrase;
- 6 phonetically rich words.

The following age distribution has been obtained: 28 speakers are below 16 years old, 129 speakers are between 16 and 30, 196 speakers are between 31 and 45, 165 speakers are between 46 and 60 and 96 are over 60.

	ELRA Members	Non Members
Price for research use	9,600 Euro	16,800 Euro
Price for commercial use	12,000 Euro	16,800 Euro

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

ELRA-S0110 Luxemburgish-German SpeechDat(II) FDB-500

The Luxembourgish-German SpeechDat(II) FDB-500 comprises 560 Luxembourgish-German speakers (247 Males, 313 Females) recorded over the Luxembourgish fixed telephone network. Each phrase or word was repeated about one time. The recording were performed by Lernout & Hauspie in Luxembourg. This database is partitioned into 3 CDs, each of which comprises 160 to 200 speakers' sessions. 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.
The following items were recorded:
- 7 application words;
- 1 sequence of 10 isolated digits;
- 4 isolated digits;
- 5 connected digits: 1 area code, 1 spontaneous phone number, 1 credit card number (15-16 digits), 1 sheet number, etc;
- 3 dates: 1 spontaneous date, e.g. birthday, 1 prompted date, 1 general and relative date expression;
- 1 embedded application word;
- 3 spelled words;
- 1 currency money amount;
- 1 natural number;
- 6 names extracted from the directory: 1 surname, 1 city of birth, 1 most frequent cities, 1 most frequent company/agency, 1 forename surname, etc;
- 4 yes/no questions;-
10 phonetically rich sentences;
- 2 time phrases: 1 time of day (spontaneous), 1 time phrase;
- 6 phonetically rich words.

The following age distribution has been obtained: 5 speakers are below 16 years old, 113 speakers are between 16 and 30, 174 speakers are between 31 and 45, 184 speakers are between 46 and 60 and 84 are over 60.

	ELRA Members	Non Members
Price for research use	9,600 Euro	16,800 Euro
Price for commercial use	12,000 Euro	16,800 Euro

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