The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

ELRA

April - June 2003

Vol.8 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.

Editor in Chief:
Khalid Choukri

Editors:
Khalid Choukri
Valérie Mapelli
Magali Jeanmaire

Layout:
Magali Jeanmaire

Contributors:
J.P. Cante
Khalid Choukri
M. Glesser
Pat Hall
L. Lelièvre
Valérie Mapelli
Bayan Abu Shawar
Audrey Zeini

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
Web sites:
http://www.elra.info or
http://www.elda.fr

Dear Colleagues,

This is the second issue of the quarterly ELRA newsletter for 2003. The Annual General Assembly of the Association that was organised in April put the focus on the 3 missions and services of particular relevance for you as member of the HLT community. This means that ELRA and its operational unit, ELDA, will continue to offer for the future the production and the validation of Language Resources, and the evaluation of Language Technologies, in addition to its traditional activity of LR distribution.

Three technical committees were established by ELRA to take care of each activity in relation with these services: a production committee, PCom; an evaluation committee, ECom; and a validation committee, VCom.

The latter has been operational for more than 2 years and steers a network of Validation Centres for Spoken Language Resources and Written Language Resources (VC_SLR, VC_WLR) which ensure the quality of the LRs distributed by ELRA and ELDA and elaborate on related methodologies. You may have noticed that some SLR have already been validated and the corresponding validation reports entry made available from their description on the on-line catalogue. This work has been carried out by the SPeech EXpertise center (SPEX), ELRA VC_SLR, located in The Netherlands, which is in charge of conducting the validation work and other Quick Quality Check processes aiming at reporting on the quality of the speech resources. Concerning the validation of the WLR, the Center for SprogTeknologi (CST), as head of a network of VC_WLR, will review the validation criteria for WLR and then start the validation work in cooperation with the VC_WLR for the different types of LRs.

Evaluation is a second major mission within which ELRA and ELDA are particularly active. The evaluation of language technologies allows to assess the performance of a given technology. Evaluation benefits both to the developers, as they will get a chance to identify the problems and find the solutions, and to the funding agencies and end-users, who will be able to measure the progress achieved and how the invested money led to the development of useful new technologies. At ELDA, a new department has been set up to handle all the activities related to the evaluation of language technologies, conducted notably in the framework of French and European evaluation campaigns.

ELRA and ELDA offer a third service of interest for the HLT community, involved either in research projects or in industrial developments, which consists in the collection and production of LRs. Recently, we have produced, or supervised the production of, a significant number of SLR in the framework of 2 European projects, Speecon (for the development of voice-driven interfaces for consumer applications) and OrienTel (for the development of multilingual interactive communication services in Mediterranean countries), in addition to projects commissioned by industrial partners. These SLR are to be included in ELDA's catalogue for their distribution to the HLT community.

During the Annual General Assembly, a new Board member was elected, in the seat left vacant by the passing of Angel Martin Municio: Pasi Tapanainen, from the Conexor Oy company in Finland, will represent the written college at the ELRA Board.

We would also like to take the opportunity to mention here 2 new Integrated Projects, which have been short listed within the 6th Framework Programme of the European Commission, in which ELRA and ELDA will be involved for the next three years. The first one is CHIL (Computer-Human Interaction in the Loop). CHIL aims at improving interactivity between the human users and the computers. Indeed, these will be expected to observe and understand the users, modelling their activities and inferring their intentions and needs. The second one is TC-Star, which aims at making Speech-to-Speech Translation (SST) become a reality. ELDA was involved in TC-Star_P, the preparatory phase of the TC-Star project. TC-Star_P was driven by industrial requirements and involved industrial key players active in the development of SST systems and components, academic research institutions active in SST systems and components research, infrastructure centres active in the development of LRs for SST components and SMEs using the provided technologies. ELDA will work on the aspects related to LRs and evaluation within both projects.

Furthermore, ELRA and ELDA are in charge of the organisation of 2 major events in the area of HLT, with the collaboration of French and European partners: LangTech 2003 and LREC 2004.

LangTech 2003 is the 2nd edition of the European forum for language technology and will be held in Paris on 24th and 25th November 2003. LangTech 2003 is turned towards the commercial aspects of existing speech and language technologies and is definitely business-oriented. It will combine 3 parallel conference sessions, in the areas of voice technologies and applications, semantic web and knowledge management and multilinguality; an exhibition, where companies will showcase their products and services; and 2 "elevator pitch" sessions dedicated to SMEs, which will get a 5-minute time slot to present their activities to attract customers and investors.

The 4th edition of the Language Resources and Evaluation Conference, organised by ELRA since 1998, will take place next year in Lisbon, Portugal, from 24th to 30th May 2004. LREC 2004 first announcement is available at the end of this issue.

In addition to the 1st LREC 2004 announcement, you will find in this newsletter a review of the workshop on "Computational Linguistics for the Languages of South Asia" which was organised within the EACL 2003 conference. This review was prepared by Pat Hall, in the framework of the SCALLA (Sharing CApabiLity in Localisation and human LAnguage Technologies) project. The second article was written by Bayan Abu Shawar after the CL 2003 conference, and offers a summary of some papers which were presented there dealing with e.g. corpus processing, etc. A third paper is presented in this issue about the European ANITA (Audio eNhancement In Telecom Applications) project, which aims at reducing audio acoustics noise in secured communications. This article was submitted by EADS Telecom, a partner of the ANITA consortium. A 4th article by Valérie Mapelli and Khalid Choukri elaborates on the ELARK initiative, standing for Extended LAnguage Resource Kit, and a follow-up of BLARK (Basic Language Resources Kit). It was initially prepared for the ENABLER (European National Activities for Basic Language Resources) network.

Last but not least, the new resources added to our catalogue are listed at the end of this newsletter.

We wish you a Happy Holiday!

Joseph Mariani, President

Khalid Choukri, CEO

Computational Linguistics for the Languages of South Asia

Pat Hall

The SCALLA project ran a workshop at the 2003 meeting of the European Association of Computational Linguistics. This was a formal workshop of the EACL 2003 conference, titled as above, and subtitled "Expanding Synergies with Europe". We had been encouraged to organise a workshop by one of the organisers of EACL2003, and went through the normal processes of soliciting papers, reviewing these, resulting in the accepted papers as listed at the end.

The keynote presentation "Creating Language Resource for NLP in South Asian Languages" was given by Professor Rajeev Sangal from the Language Technologies Research Centre at the International Institute of Information Technology in Hyderabad, India. This led into a series of sessions where people presented their papers, or other related work, with lots of stimulating discussion. The following summarises some of the highlights of the meeting.

The languages

The languages of South Asia are interesting linguistically, but are also important economically, with the population of the region approaching one and a half billion, and with migrant communities across the world, including significant numbers in Europe, particularly in the United Kingdom.

English is widely understood throughout South Asia, and as a visitor to South Asia you may well gain the impression of a greater competence than in fact there is - only some 5% of the population could be deemed literate in English. Local languages are still widely spoken: the range of languages of the area was discussed by both Hussain for Pakistan and Mallikarjun for India. There are some 17 official languages in India, with another 5 official languages elsewhere in the region - the map below shows just some of these and their geographical distribution. Note that several of these are spoken across national boundaries, and that Tamil is also an official language of Singapore. But very many other languages are widely used in the region, with over 500 distinct languages recognised; of course the impreciseness of the distinction between a language and dialect makes it impossible to give a precise number - some accounts put the figure as high as 1652 languages in India alone, based on the 1961 census, quoted by Mallikarjun. Many of these languages are only spoken by a few thousand persons and must be viewed as endangered.

The main thing to note is the large number of distinct languages, some of which were traditionally written, but very many of which are either unwritten or have only recently acquired writing systems. The new writing systems themselves would be based upon the system of the local dominant language, and thus be an Arabic system in Pakistan and a Brahmi system elsewhere in South Asia. Even something as seemingly simple as writing can be problematic, as was illustrated by the paper by Sojka and Antoš for Thai which is written in a Brahmi writing system.

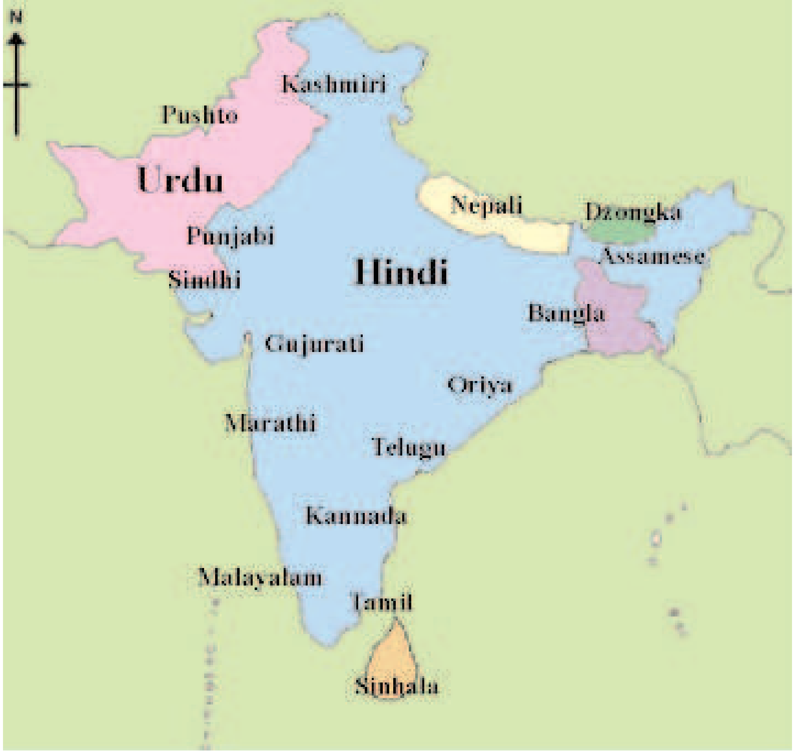
The languages themselves mostly belong to three major groupings - Tibeto-Burmese in the northern Himalayan region including Dzongka and Tibetan; Indo-European in the northern central regions including Urdu and Hindi, Bangla, Punjabi, Nepali, and many others including Sinhala from Sri Lanka; and Dravidian in the south including Tamil, Kannada, Telugu and Malayalam. The languages of scholarship have been Sanskrit, more recently Farsi, and very recently English.

Much more needs to be learnt about these languages, even the major languages, through direct collection of empirical data. This is the role of corpus linguistics, covered in the workshop by Mallikarjun and Baker et al. McEnery in presenting the Baker et al paper distributed a CD containing corpora gathered by the EMILLE project.

The computational linguistics.

To be able to process the languages of South Asia we need good lexicons and a sound understanding of the structure of the language. Sangal in his keynote gave an excellent survey of the current state of computational linguistics in South Asia through the medium of machine translation which demands such a full spectrum of capabilities (we have no written paper, but see http://www.iiit.net/ltrc/Publications/Techreports/toc.html for related reports). Within the context of much that has already been achieved, he highlighted many problems that still need to be solved. This complemented well the survey by Hussain of computational linguistics in Pakistan.

Several papers in the workshop looked at problems of computational linguistics, sometimes with more focused applications in mind. One issue that emerged in the discussions around these was the need to exploit structures discovered in one language in understanding other closely related languages. It was agreed that nowhere was this more necessary and appropriate than between Urdu and Hindi, very closely related languages (see Kachru 1987) but which seem so superficially different, the one being written in an Arabic derived system and the other in Devanagari derived



from Brahmi. So for example, the complex predicates described by Butt et al would also be applicable to Hindi and other closely related languages like the Oriya discussed by Shabadi.

We also saw some interest in universal approaches to language. Mukerjee in presenting the paper by Goyal et al emphasised the promise of UNL, the Universal Network Language, form the United Nations. Butt also talked at length about Lexical-Functional Grammars, which she and colleagues applied to Urdu but are also applying to a wide range of other languages. However this stirred memories of interlingua and a concern that the pursuit of universal solutions might be equally doomed to failure.

Speech is important in South Asia and more generally throughout the developing world, particularly for non-literate peoples whose only means of access to information on the internet is through speech. Shalonova and Tucker are looking to Text-To-Speech (TTS) systems for South Asian languages, looking to exploit commonalities across all spoken languages. While a universalist approach to languages might have difficulties, it was felt that a universal approach might work for speech.

The future of SCALLA

The SCALLA project is now in its final year. It has one more conference to organise, this time back in South Asia. The final conference will be held in Kathmandu, Nepal, in December 2003 or January 2004, and will focus on how Natural Language Processing can help the developing world through overcoming barriers to access to IT and modern technology.

Workshop papers

These papers have been published by EACL as a Workshop Proceedings. Copies can be obtained via the workshop website at <http://computing.open.ac.uk/Sites/EACLSouthAsia>.

Baker, Paul; Andrew Hardie, Tony McEnery and B.D. Jayaram, Corpus Data for South Asian Language Processing

Butt, Miriam; Tracy Holloway King and John T. Maxwell III, Productive Encoding of Urdu Complex Predicates in the ParGram Project

Goyal, P; Manav R. Mital, A. Mukerjkee, Achla M. Raina, D. Sharma, P. Shukla and K. Vikram, Saarthak: A Bilingual Parser for Hindi, English and Code-switching Structures

Gupta, Deepa; and Niladri Chatterjee, A Morpho-Syntax Based Adaptation and Retrieval Scheme for English to Hindi EBMT

Hussain, Sarmad, Computational Linguistics (CL) in Pakistan: Issues and Proposals

Mallikarjun, B, Corpora in Minor Languages of India: Some Issues

Ramanathan, Ananthakrishnan; and Durgesh D. Rao, A Lightweight Stemmer for Hindi

Shabadi, Kalyani R., Finite State Morphological Processing of Oriya Verbal Forms

Shalonova, Ksenia; and Roger Tucker, South Asian Languages in Multilingual TTS-related Database

Sojka, Petr; and David Antoš, Context Sensitive Pattern Based Segmentation: A Thai Challenge

References

Kachru, Yamuna, Hindi-Urdu, Chapter 3 in The Major Languages of South Asia and the Mikdels East and Africa, edited by Bernard Comrie. Routledge 1987.

Pr. Pat Hall
Open University
Computing Department
MK7 6AA Milton Keynes
United Kingdom
Tel: +44 (0)1908 652694
E-mail: p.a.v.hall@btinternet.com

Review of CL2003, the International Conference on Corpus Linguistics

Bayan Abu Shawar

The first International Conference on Corpus Linguistics was organised in 2001 (Rayson et al 2001) to honour Geoffrey Leech on his 65th birthday. CL2003 was a repeat by popular demand. It was held at Lancaster University from 28th to 31st March 2003, and gathered 210 participants.

The conference included four one-day workshops, and four days of main papers sessions; one pre-conference workshop produced separate Proceedings (Simov and Osenova 2003), other papers were included in the main conference Proceedings (Archer et al 2003). Talks were presented in three different parallel sessions, classified according to topics such as: corpus building, parsing techniques, translation studies, corpus tools, taggers, morphosyntactic annotation, information extraction, semantics, text comparison, European minority languages, word frequency studies, Internet English, and speech annotation. A wide range of domains related to corpus-based natural language processing was covered.

Five invited talks were presented: "What can corpus linguistics tell us about linguistic creativity?" by Michael Hoey, "Everything you wanted to know about the American National Corpus but weren't afraid to ask" by Nancy Ide, "Frame, phrase, or function: a comparison between frame semantics and local grammars" by Susan Hunston, "Are we nearly there yet, mum?" by Geoffrey Sampson, and "Corpora and the lexicon" by Nicoletta Calzolari.

These five talks were exciting, discussing new issues and ideas related to language and corpora.

You will find in this short review a summary of some papers of particular interest for our research work in Narural Language Processing, which consists of developing machine-learning techniques to train a chatbot using a corpus-based approach, implementing software that convert the readable text, namely the corpus, into the chatbot language model format.

Stefan Grondelaers presented " a corpus-based approach to informality: the case of Internet chat".

He described the language of the Internet Relay Chat (IRC) as an example of "spoken language in written form", as it shares the informality characteristic with the spoken language. He classifies the informality characteristics into four types. The first is the dialogue character represented in the higher frequency of 2nd person pronouns and vocatives. The second is that a lot of abbreviation and ellipses are typed by the users. A third source is speaker related: age, gender, and the topic of chatting. The fourth factor is register-related: chatters might choose to sound colloquial, or to maintain a more formal standard. Then he illustrated two methods to identify the informality, stylometric approaches and the onomasiological profile. Stylometric approaches are based on the calculation of isolated variables, used to identify the first three types of informality characteristics. Onomasiological profiles were used to compare lexical, morphological, syntactic and phonological preferences in Belgian IRC logs and other modes of written communication. The calculations revealed that the four types of informality do not coincide in Belgian Dutch IRC.

He concluded that the linguistic specificity of chat is determined in the first place by the production demands (speed and turn-taking efficiency) of spoken conversation.

In his paper, Jun Takahashi examined interaction via computer mediated communication (CMC): "Do we talk or write differently over the net?".

He reviewed previous work that described the characteristics of the language of CMC and the relationship between CMC and spoken and written language. Theses characteristics describe CMC as dynamic communication, classified into two types: synchronous such as IRC, which maintains strong coherence in turn-taking, like face-to-face conversation; and asynchronous such as emails and Bulletin Board Services (BBS). CMC looks like writing, as users enter manually their discourse, deliver textual information, but the language is informal: a lot of pronouns and modal auxiliaries arise as in spoken language.

He investigated English as a multi-national language (EML) in the Internet discussion forum based on CMC. He gathered data from different corpora: the NET-EN Corpus (NC), which consists of discussions in e.g. the political, social, hobbies and sport domains, the Base text Corpus (BC) and the Response text Corpus (RC). All these corpora show EML in use by Japanese English users.

Then he illustrated the analysis method he used by comparing the 50 most frequent lexical items between written, spoken, interviews & letters (Int&Ltt) English from the native-speaker British National Corpus (BNC) and Net_EN corpus (NC). Another aim was to find words with unique uses and investigate these in terms of sense and patterns of use. Finally he concluded that this study shows the unique use of English produced by Japanese users of English as multi-national language (EML) in today's context, computer mediated communication (CMC).

Another interesting paper was entitled "Relating lexical items to sociolinguistic features in a spontaneous speech corpus of Spanish" by Jose Maria Guirao Miras et al. They presented the correlation between linguistic and socio-contextual features using the C-ORAL-ROM corpora, a multilingual corpus of spontaneous speech for four basic Romance languages, namely French, Italian, Portuguese and Spanish. Nine participants, including ELDA, are working on this project, funded by the EU. The main goal of C-ORAL-ROM is to compare the four languages on the same grounds and provide comparative studies at different linguistic levels. In order to do the comparison, each subcorpus followed the same text distribution (sampling design and text size) and standard format (using XML to denote header and transcript). New computational tools give you an analysis of the linguistic features of each sub-corpus at three levels: word, lemma and POS, using log-likelihood ratio. At the POS level, every word is related with the speaker and the text, every socio-contextual information is stored in the file header and, depending on the different header features, many sub-corpora can be derived from the corpus, for example a male sub-corpus, a telephone sub-corpus, etc.

They illustrated a new idea based on an n-grams algorithm which was used at word level to extract multi-words candidates. All the n-grams with three or more occurrences, for n=4, 3, and 2 were removed. Then a filter process was applied that discarded all n-grams which started or ended with determiner or auxiliary. Finally, multi-words were selected by hand and each one was treated as a single lexical unit. They showed some results of this tool. For example, according to POS analysis, men prefer to use nouns while women prefer pronouns at the POS level, with respect to the Spanish corpus.

The intention is to compare the four romance languages after finishing the morphosyntactic annotation. This work will be published as a chapter in a book "Corpus Linguistics around the world".

Prof. Geoffrey Leech and Martin Weisser presented the SPAAC tool within their paper "Generic speech act annotation for task-oriented dialogues".

SPAAC (Speech Act Annotated Corpus) is the XML semi-automatic tool developed to annotate dialogues in terms of speech acts. They applied this tool within a pilot project, which has attempted to achieve a middle ground between general coverage of dialogues aim and specific tasks or domains annotation aim.

Dialogue corpora were collected from two sources: first, BT OASIS dialogues which involves 100 calls to the operator and 150 calls to BT residential enquiries; second, the Train-Line dialogues which involves calls to a call centre providing railway timetable information and tickets, and seat reservation services.

They illustrated the main jobs of SPAAC as follows: automatic conversion of the text files containing the transcription to XML markup, interactive segmentation of the dialogues into utterance-units (called C-units), automatic assignment of speech-act categories, together with other categories giving information on the form (declarative, yes-no question, imperative), the polarity (positive, negative), the topic (relating to train journey location, name, day, date, time) and the mood (semantic categories such as probability, reason). The next step consisted of the manual post-editing and the correction of speech-act tags.

One of the possible uses of speech act annotated corpora is as a training corpus for the modelling of dialogue behaviour, using either statistical, rule/structure-based or hybrid language model.

The conference also had several software demo sessions. For example, Paul Rayson and Olga Moudraia demonstrated the W-matrix tool. W-matrix was implemented by Dr. Rayson to compare different sized corpora at three levels: word, POS and semantic-tagged. The comparison results are viewed as frequency lists ordered by log-likelihood ratio, indicating the most important differences between corpora.

One of the most important achievements for me and my research group lies in the fact that 10 papers from University of Leeds were presented. Eight students and postdocs of my group (Natural Language Processing) presented papers, in addition to our supervisor Eric Atwell who presented two papers.

We have collected extended abstracts in a separate research report on "Corpus Linguistics, Machine Learning and Evaluation: views from Leeds" (Atwell et al 2003). This report includes: "Using Dialogue Corpora to Train a Chatbot" (Bayan Abu Shawar, Eric Atwell); "A Word-Token-Based Machine Learning Algorithm for Neoposy: coining new Parts of Speech" (Eric Atwell); "Detecting Student Copying in a Corpus of Science Laboratory Reports: Simple and Smart Approaches" (Eric Atwell, Paul Gent, Julia Medori, Clive Souter); "Statistical modeling of MT output corpora for Information Extraction" (Bogdan Babych, Anthony Hartley, Eric Atwell); "Rationale for a Multilingual Aligned Corpus for Machine Translation Evaluation" (Debbie Elliott, Anthony Hartley, Eric Atwell); "The Human Language Chorus Corpus HULCC" (John Elliott, Debbie Elliott); "A survey of Machine Learning approaches to analysis of large corpora" (Xunlei Rose Hu, Eric Atwell); "Using the HTK speech recogniser to analyse prosody in a corpus of German spoken learners' English" (Toshifumi Oba, Eric Atwell); "The Use of Corpora for Automatic Evaluation of Grammar Inference Systems" (Andrew Roberts, Eric Atwell); and "Methods and tools for development of the Russian Reference Corpus" (Serge Sharoff).

Belinda Maia and Luis Sarmento won the prize for the best poster design, on "Constructing comparable and parallel corpora for terminology extraction"

Overall, the CL2003 conference was a nice opportunity to see the work of other people in different countries and to chat directly with them. Participants enjoyed the friendly, familiar environment, the delicious food and professional organization. Many thanks to: Tony McEnery, Dawn Archer, Paul Rayson, Andrew Wilson, Paul Baker, and all assistants from Lancaster University who organized the CL2003.

References

Eric Atwell, Bayan Abu Shawar, Bogdan Babych, Debra Elliott, John Elliott, Paul Gent, Anthony Hartley, Rose Hu, Julia Medori, Toshifumi Oba, Andrew Roberts, Serge Scharoff, Clive Souter. 2003. Corpus Linguistics, Machine Learning and Evaluation: Views from Leeds. Research report 2003.02, School of Computing, University of Leeds. http://www.comp.leeds.ac.uk/research/pubs/reports.shtml

Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.) 2003. Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16. UCREL, Lancaster University. http://www.comp.lancs.ac.uk/ucrel/tech_papers.html

Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja (eds) 2001. Proceedings of the Corpus Linguistics 2001 conference. UCREL technical paper number 13. UCREL, Lancaster University

Kiril Simov and Petya Osenova (eds.) 2003. Proceedings of The Workshop on Shallow Processing of Large Corpora (SProLaC 2003) held in conjunction with the Corpus Linguistics 2003 conference. UCREL technical paper number 17. UCREL, Lancaster University.

Miss Bayan Abu Shawar
University of Leeds
School of Computing
Computer Vision and Language
Research Group
LS2 9JT United Kingdom
http://www.comp.leeds.ac.uk/bshawar

ANITA, Audio eNhancement In Telecom Applications

J.P. Cante, M. Glesser, L. Lelièvre and A. Zeini

ANITA (Audio eNhancement In Telecom Applications) is a European project launched on the initiative of EADS TELECOM with the objective of reducing audio acoustics noise in secured communications.

Context and objectives of the study

ANITA is a R&T project funded within the 5th RTD framework program of the European Commission under the Information Society Technology programme. This project started in April 2002. It will run for 30 months and come to an end in October 2004.

The ANITA consortium has been set up to answer one of the strongest needs of secured telecommunications users like fire brigades, police and health services that are not met today by existing products. Its main objective is to reduce audio acoustics noise in secured communications in adverse environments (sirens, alarms, engines, water pumps, stress situations, etc.)

Two solutions will be developed within ANITA: the 1st one for handheld mode applications and the other for in-car applications. For this purpose, single channel and multichannel (based on an array of 8 microphones) algorithms have been developed.

In Europe, the secured digital telecommunications are based on two digital standards: TETRA and TETRAPOL.

- TETRA (Trans European Trunk Radio), ETSI standard (European, Telecommunications Standard Institute), is supported by 65 industrials in 16 countries.
- TETRAPOL, recognised by ITU (International Telecom Union) is supported by more than 30 industrials including EADS TELECOM.

ANITA will comply with both standards.

ANITA is structured as follows:

- Project management
- Definition of end-users requirements
- Production of an audio database including both voice and noise recordings
- Multichannel approach based on a microphones array
- Single-channel approach
- Integration of algorithms and prototyping
- In-site validation
- Dissemination and exploitation.

Partners

- EADS TELECOM, as project leader, is a worldwide provider of secured telecommunications networks for defence, civil secured and public safety markets. EADS TELECOM has developed a database including voice and noise recordings and will participate in the prototypes manufacturing and the on-site validation
- FULCRUM Voice Technologies, an English manufacturer of noise reduction solutions and expert in voice technology and related products, will be in charge of the prototypes manufacturing
- KTH (Kungl Tekniska Högskolan, Royal Institute of Technology), a Swedish university, will develop noise reduction algorithms based on a single-channel approach
- FAU (Friedrich Alexander University), a German university, will develop noise reduction algorithms based on a multichannel approach
- TRADIA, a Spanish TETRA/TETRAPOL operator, is responsible for the end-users requirements definition and is in charge of the prototypes validation.

Description

End-user requirements definition and database recordings including both single-channel and multichannel recordings

In order to assess performances of current equipments used by some emergency organisations, a study has been made upon the current situation of public safety networks from both technical and subjective audio quality points of view: questionnaires were submitted to 3 Spanish organisations (fire brigades, police and health services) using TETRA and TETRAPOL networks.

EADS TELECOM has collected a database made up of three types of recordings:

- Multi-channel recordings,
- Single-channel voice recordings,
- Single-channel noise recordings.

Multi-channel recordings were carried out in Barcelona, in a patrol car and a fire brigade truck.

A linear uniformly-spaced 8-element array was installed above the passenger's head and the same array was available between the driver and the passenger, to enable the driver to communicate with the remote talker.

The microphone array was mounted on the ceiling of both vehicles so that the distance between the microphones and the speaker reached approximately 0.5 meters.

The recordings have been made in English, following different scenarios: car stopped and normal driving, with closed and open windows, with a speaker and without speaker, with siren and without siren.

Single-channel recordings include 41 voices in English, French, German and Spanish.

The speaker could read the figures and words on a computer screen with a close microphone.

The database includes male and female voices under calm, stressed and in panic conditions. It includes:

- Recorded speech data under calm conditions are made of:
  - 60 phonetically rich sentences provided by ELDA in 4 languages (about 8 minutes of recording).
  - Figures and words (around 8 minutes).
  - 10 minutes text.
- The recorded speech data under stressed and in panic conditions are made of:
  - The same 60 phonetically rich sentences
  - The same figures and words.

These recordings were carried out EADS TELECOM premises in Bois d'Arcy at the end of 2002/beginning of 2003 with the participation of bilingual employees, ANITA partners and five European comedians for stressed and in panic conditions.

Out of 41 speakers, 71% were native people (English, French, German and Spanish).

The database also contains noise recordings:

- In-car/pedestrian sound recordings have been carried out in real traffic around EADS TELECOM premises in Montigny-le-Bretonneux. The car (Ford Mondeo) was equipped with a siren.
- In-car recordings (Opel Corsa) without siren have been carried out in Paris (main streets and highway).
- Outside recordings (e.g. public work noise or traffic on a crossroad) have been carried out in Paris (main streets and highway).
- Public transport recordings have been carried out on Paris public transport network (in train and subway).

Multichannel recordings (voice and noise recordings) have also been carried out in Barcelona by FAU, TRADIA and EADS TELECOM, in July 2002. The recordings took place at the Fire Brigade Headquarter station. A uniformly spaced linear microphone array with 8 cardioid microphones was used.

Algorithms development for handset and in-car applications

These recordings (voice and noise) will form a test base for the Swedish and German universities in charge of developing the noise reduction algorithms. Two types of algorithms are being developed since July 2002:

- Algorithms based on a single-channel approach, which exploits in a first phase detailed models based on a-priori knowledge of both noise and speech signals. This type of algorithms addresses handheld mode applications
- Algorithms based on a multichannel approach (use of a microphone array with 8 sensors): robust adaptative beamforming algorithms will be optimised and acoustic echo cancellation will be included. Algorithms for blind source separation will be investigated. This type of algorithms addresses in-car applications.

Integration and prototyping

These algorithms will then be implemented into TETRA/TETRAPOL prototypes by FULCRUM Voice Technologies, specialised in signal processing. The prototyping task began in April 2003.

In-site validation of prototypes

The Spanish operator TRADIA, supported by EADS TELECOM, will carry out the "field" tests in real life situations. It will start in October 2003, for 1 year. Voice quality and voice intelligibility obtained with the ANITA prototypes will be assessed over TETRA and TETRAPOL networks (Agora and Nexus networks operated by TRADIA in Catalonia) and will be compared to the initial voice quality and intelligibillity.

ANITA Web site

An Internet site devoted exclusively to this project is at your disposal at the following address: http:// anita.eads-telecom.com/anita/

Contact

If you would like to obtain more information, you should contact Audrey Zeini at audrey.zeini@eads-telecom.com.

References

[1] M. Kuropatwinski, and W.B. Kleijn, "Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding", Proc. ICASSP, 2001.

[2] O. Hoshuyama, A. Sugiyama and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," IEEE Trans. on Signal Processing, vol. 47, no. 10, pp. 2677-2684, October, 1999.

[3] W. Herbordt, H. Buchner, W. Kellermann, "An acoustic human-machine front-end for multimedia applications," European Journal on Applied Signal Processing, pp. 1-11, January, 2003.

[4] W. Herbordt, W. Kellermann, "Adaptive beamforming for audio signal acquistion," in J. Benesty, Y. Huang (eds.),"Adaptive Signal Processing - Applications to Real-World Problems," pp. 155-194, Springer-Verlag, Berlín, Germany, 2003.

[5] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. on Signal Processing, pp. 504-512, Julio, 2001.

[6] I. McCowan, H. Bourlard, "Microphone array post-filter for diffuse noise field," Proc. ICASSP'02, vol. 1, pp. 905-908, May, 2002.

[7] D. Florencio, H. Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays," Proc. ICASSP, vol. 1, pp. 197-200, May, 2001.

[8] J. J. Shynk, "Frequency-domain multirate adaptive filtering," IEEE Signal Processing Mag., vol. 9, no. 1, pp 14-37, January, 1992.

J.P. Cante, M. Glesser, L. Lelièvre,
A. Zeini
EADS TELECOM
Rue J.P. Timbaud
Montigny Le Bretonneux
78063 Saint Quentin Yvelines Cedex
France
E-mail: audrey.zeini@eads-telecom.com
http:// anita.eads-telecom.com/anita

Towards an Extended LAnguage Resource Kit (ELARK): summary of ELDA's report within the ENABLER thematic network

Valérie Mapelli and Khalid Choukri

In the framework of the ENABLER thematic network (European National Activities for Basic Language Resources - Action Line: IST-2000-3.5.1), ELDA elaborated on a report defining a (minimal) set of LRs to be made available for as many languages as possible [MAPELLI & CHOUKRI 2003] and map the actual gaps which should be filled in order to meet the needs of the HLT field. The report aimed at providing the basics on a larger initiative in order to determine the BLARK concept more specifically. This article summarises the main elements of this report.

The BLARK Concept

The BLARK concept (Basic LAnguage Resource Kit) was first launched in The Netherlands. In his article [KRAUWER 1998], Steven Krauwer proposed a cooperative initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association) to be submitted to the Fifth Framework Programme of the European Commission. Due to time constraints, such a proposal was not submitted to FP5 but the concept has been adopted and popularised by many players. In particular, an initiative, adopting the BLARK concept, was launched for the Dutch language.

A Dutch initiative, called Dutch Human Language Technologies Platform was initiated in April 1999 by the Dutch Language Union (Nederlandse Taalunie), a Dutch/Flemish intergovernmental organisation responsible for strengthening the position of the Dutch language, further to an exploratory survey on the position of the Dutch in language and speech technology developments carried out between October 1997 and June 1998. This initiative aimed at stimulating collaboration between all actors involved and co-operation between Flanders and the Netherlands, and also at encouraging Flemish and Dutch participation in European projects and initiatives [CUCCHIARINI et al. 2001a] and [CUCCHIARINI et al. 2001b].

Towards an ELARK Concept

Further to their own initiative, without mentioning it as a "BLARK" initiative, a number of organisations, in particular ELRA (European Language Resources Association), ELSNET (European Network of Excellence in Language and Speech), and the LDC (Linguistic Data Consortium), have contributed to the identification, promotion, dissemination, production, etc. of Language Resources and related tools as a support to the HLT community.

Further to the definition of BLARK which focussed on the LRs needed for each language, one may face other existing and more sophisticated tools and systems that are also capable of processing language data. For instance, basic tools may cover lemmatisation, tokenisation, morphological analysis, parsers, speech analysis (front-ends), acoustic modelling, language modelling, etc., while sophisticated applications may cover information/document retrieval (spelling/grammar checkers), machine translation, named entity recognition, speech transcription, speech synthesis, etc. In many cases, sophisticated tools are a combination of many basic tools that require BLARKs to be developed. In many other cases, such sophisticated tools require extended data of their own. For instance, one may think that a speech recognition system requires a small database of isolated words to implement a basic discrete word recogniser. It is more crucial to design and train a speech recogniser[1] that transcribes audio data from broadcast news programmes. A distinction could be then made between several levels of a Language Resource Kit, the first level being a basic language resource kit "BLARK", and the other levels could be referred to as Extended LAnguage Resource Kits or "ELARK".

Table 1: Abstract from the *LRs vs. Languages* matrix

Speech Resources	fre-fr	Fre-be	Fre-sz	fre-lu	Fre-ca	fre-int	eng-gb	eng-us	eng-int	ger-de	ger-at	ger-lu	ger-int	ita-it
Broadcast speech							E	e, t	E	E				E
Articulatory database	E				E		E			E				E
Microphone/desktop speech	E		E				E	e	E	E			E	E
Read newspaper texts	E									E				E
Telephone speech database	E	E	E	E	E		E	E	E	E		E		E
Mobile-radio speech								e						
Pronunciation lexicon	E							e		E				E
Onomasticon	E						E	e		E				e
Speaker identification speech corpus			E					e						E

Legend
E=> available through ELRA,
e => exists,
“blank” => not identified/does not exist,
t => transcribed

ELRA BLARK Matrices

In the line of the promotion of the BLARK concept, a number of initiatives aimed at identifying the gaps to be filled in the HLT field. More specifically, ELRA has been working at implementing a BLARK matrix, to highlight the gaps with regards to LRs needed for specific applications and for as many languages as possible. Further to its own experience and other reports from partners such as the Dutch initiative, ELRA implemented and improved its original matrix [CHOUKRI et al. 1999] which first attempted to cross-link the types of language resources with respect to the languages that could be identified as required languages.

The abstract from the matrix Table 1 illustrates that many basic resources (as defined by ELRA) are not available for distribution or do not exist at all. In order to understand the needs in a clearer and more complete way, ELRA has extended its matrix to a list of potential applications to be cross-linked with the LRs needed and corresponding languages. This list of applications results in part from works carried out by ELRA for the French Ministries [MARQUOIS & MAPELLI 1997]. These applications are classified as follows:

1. Entering and acquiring information
- 1.1 Typing (keyboard)
- 1.2 Digitization
- 1.3 Optical Character Recognition
  - 1.3.1 Of printed characters
  - 1.3.2 Of written characters
- 1.4 Voice dictation

2. Production and documents
- 2.1 Automatic generation (words, sentences, texts)
- 2.2 Automatic generation of multimedia documents
- 2.3 Machine translation
- 2.4 Computer assisted translation
- 2.5 Voice translation
- 2.6 Speech-to-speech translation
- 2.7 Assisted localisation
- 2.8 Translation aids
- 2.9 Automatic detection and correction of errors
- 2.10 Lexical prediction
- 2.11 Advanced word processing
- 2.12 Editing aids
- 2.13 Voice commands for editing
- 2.14 Voice commands for document production
- 2.15 Automatic summarisation

3. Document management (storing, analysing and indexing)
- 3.1 Automatic indexing
- 3.2 Computer assisted indexing
- 3.3 Content analysis
- 3.4 Terminology management
- 3.5 Data compression

4. Information retrieval and presentation
- 4.1 Information retrieval
- 4.2 Help for information retrieval
- 4.3 Help for query
- 4.4 Information screening

[1] *We may take examples from other domains such as machine translation where it is mandatory to have some tools that analyse and transfer/translate one language into another with a small bilingual lexicon while a finalised/packaged product which would require a huge lexicon and would not be part of BLARK but rather part of ELARK.*

- 4.5 Information analysis and selection
  - 4.5.1 Mapping information
  - 4.5.2 Relevance
- 4.6 Automatic summary
- 4.7 Synthesis
- 4.8 Navigation

5. Information dissemination

- 5.1 Information servers
- 5.2 Routing information
  - 5.2.1 Calls and switchboard
  - 5.2.2 Workflow
  - 5.2.3 Voice and electronic mailing
- 5.3 Selective dissemination of information (SDI)
- 5.4 Electronic data interchange (EDI)

6. Securing information access

- 6.1 Information privacy
- 6.2 Identification and verification of the user and of the origin of the data
- 6.3 Information integrity

An abstract of the resulting matrix (Production of documents) is given in Table 2[2].

The two BLARK matrices as proposed by ELRA aim to be cross-linked, made accessible and modifiable directly from the ELRA web site. This will enable external customers or providers of LRs to fill it in with complementary information and help ELRA at identifying available LRs and promoting the production of new specific ones. At a first step, the combined matrices will be submitted to experts of the HLT field for validation. This could be done through an extended survey and/or the implementation of the matrix online through the ELRA web site.

In a near future, any customer or LR provider aware of an existing LR will be able to complete the cross-linked matrices, pointing to an existing LR. This information will be then considered directly at ELDA in order to check the accuracy of the information. When this information is confirmed, the corresponding cells in the matrix will be filled in accordingly and made available online.

Table 2: Abstract from the *Applications vs. Types of LRs* matrix

	Speech Resources	Broadcast speech	Articulatory database	Microphone/desktop	Read newspaper texts	Telephone speech	Mobile-radio speech	Pronunciation lexicon	Onomasticon	Speaker identification	**Lexica**	Monolingual lexicon	Multilingual lexicon	**Text Corpora**	Broadcast text corpus	Conversation text	Newswire text corpus	Monolingual corpus	Multilingual and parallel	Treebank
2 Production of documents																				
2.1 Automatic generation (words, sentences, texts)												X	X					X	X	X
2.2 Automatic generation of multimedia documents			X	X				X	X			X	X					X	X	X
2.3 Machine translation												X	X		X	X	X	X	X	X
2.4 Computer assisted translation												X	X		X	X	X	X	X	X
2.5 Voice translation			X					X	X										X	X
2.6 Speech to speech translation		X	X	X		X		X	X							X				

Conclusion

In the future, such an initiative, combined with all ongoing initiatives (and hopefully many more) focussing on the same goal, should contribute to map and, in the end, fill, if not all, at least a fair number of the gaps that should improve the working material of the HLT community. In these initiatives, we should not omit the maintenance work on Language Resources, further to the production work, as was raised in [MACLEOD 1998]. In her article Catherine Macleod proposed that "along with the mandate and the funding to create a resource, thought should be given to how and at what level the resource should be supported". Indeed, expenses on LRs are big enough to take into consideration their reusability on a long-term, therefore maintenance and updating are rather important issues.

Bibliography

[CHOUKRI et al. 1999] Khalid Choukri, Valérie Mapelli and Jeff Allen, New Developments within the European Language Resources Association (ELRA), in Proccedings EUROSPEECH 1999.

[CUCCHIARINI et al. 2001a] Catia Cucchiarini, Walter Daelemans and Helmer Strik, Strengthening the Dutch Human Language Technology Infrastructure, in ELRA Newsletter Vol. 6 N. 4. 2001.

[CUCCHIARINI et al. 2001b] Catia Cucchiarini, Walter Daelemans and Helmer Strik, Strengthening the Dutch Language and Speech Technology Infrastructure, in Proceedings COCOSDA 2001.

[KRAUWER 1998] Steven Krauwer, ELSNET and ELRA: A common past and a common future, in ELRA Newsletter Vol. 3 N. 2. 1998.

[MACLEOD 1998] Catherine Macleod, A Plea for Consideration of Maintenance of Language Resources, in Proceeding LREC 1998.

[MAPELLI & CHOUKRI 2003] Valérie Mapelli, Khalid Choukri, Deliverable 5.1 Report on a (minimal) set of LRs to be made available for as many languages as possible, and map of the actual gaps, internal report, 2003 (http://www.enabler-network.org/reports.htm).

[MARQUOIS & MAPELLI 1997] Emilie Marquois, Valérie Mapelli, Intégration des outils linguistiques dans des systèmes de traitement de l'information professionnelle, internal report, 1997.

The complete report can be accessed from the following address:
http://www.enabler-network.org/reports.htm
ENABLER official web site:
http://www.enabler-network.org

Valérie Mapelli & Khalid Choukri
ELRA/ELDA
55-57 rue Brillat-Savarin, F-75013 Paris
E-mail: [mapelli;choukri]@elda.fr

SEPLN 2003
19th Conference of the Spanish Society for Natural Language Processing (SEPLN)
10 - 12 September, 2003. University of Alcala, Alcalá de Henares (Madrid, Spain)

Organised by the SEPLN and the Office for Spanish in the Information Society (OESI) at Instituto Cervantes

Objectives

As in previous editions, the aim of SEPLN for this Conference is to promote the dissemination of research, development and innovation activities conducted by Spanish and foreign researchers in the field of Natural Language Processing (NLP). The conference will provide a forum for discussion and communication to facilitate an effective exchange of knowledge and scientific materials that are necessary for promoting the publication of relevant work and the establishment of means of collaboration with national and international Institutions that are active in this field.

For more information about the conference: http://oesi.cervantes.es/sepln

[2] *This matrix is being updated to consider more applications and more resources.*

LREC 2004

24th-30th May 2004, Lisbon (Portugal)

www.lrec-conf.org

lrec@elda.fr

Conference Aim

The aim of this conference is to provide an overview of the state-of-the-art, discuss problems and opportunities, exchange information regarding LRs, their applications, ongoing and planned activities, industrial uses and needs, requirements coming from the new e-society, both with respect to policy issues and to technological and organisational ones.
LREC will also elaborate on evaluation methodologies and tools, explore the different trends and promote initiatives for international collaboration in the areas covered by Human Language Technologies (HLT).

Conference Topics

Issues in the design, construction and use of Language Resources (LRs)

- Guidelines, standards, specifications, models and best practices for LRs,
- Methods, tools and procedures for the acquisition, creation, management, access, distribution and use of LRs,
- Methods for the extraction and acquisition of knowledge (e.g. terms, lexical information, language modelling) from LRs,
- Organisational and legal issues in the construction, distribution, access and use of LRs,
- Availability and use of generic vs. task/domain specific LRs,
- Definition and requirements for a Basic and Extended LAnguage Resource Kit (BLARK, ELARK) for all languages,
- Monolingual and multilingual LRs,
- Multimedia and multimodal LRs. - Integration of various media and modalities in LRs (speech, vision, language),
- Documentation and archiving of languages, including minority and endangered languages,
- Ontologies and knowledge representation,
- Terminology, term extraction, domain-specific dictionaries,
- LRs for linguistic research in human-machine communication,
- Exploitation of LRs in different types of applications (information extraction, information retrieval, speech dictation, translation, summarisation, web services, semantic web, etc.),
- Exploitation of LRs in different types of interfaces (dialog systems, natural language and multimodal/multisensorial interactions, etc.)
- Industrial LRs requirements, user needs and community's response,
- Industrial production of LRs,
- Industrial use of LRs,
- Metadata descriptions of LRs.

Issues in Human Language Technologies (HLT) evaluation

- Evaluation, validation, quality assurance of LRs,
- Evaluation methodologies, protocols and measures,
- Benchmarking of systems and products, resources for benchmarking and evaluation, blackbox, glassbox and diagnostic evaluation of systems,
- Usability and user experience evaluation, qualitative and perceptive evaluation,
- Evaluation in written language processing (document production and management, text retrieval, terminology extraction, message understanding, text alignment, machine translation, morphosyntactic tagging, parsing, semantic tagging, word sense disambiguation, text understanding, summarisation, question answering, localisation, etc.),
- Evaluation in spoken language processing (speech recognition and understanding, voice dictation, oral dialog, speech synthesis, speech coding, speaker and language recognition, spoken translation, etc.),
- Evaluation of multimedia document retrieval and search systems (including detection, indexing, filtering, alert, question answering, etc),
- Evaluation of multimodal systems,
- From evaluation to standardisation.

General issues

- National and international activities and projects,
- LRs and the needs/opportunities of the emerging industries,
- LRs and contributions to societal needs (e.g. e-society),
- Priorities, perspectives, strategies in national and international policies for LRs,
- Needs, possibilities, forms, initiatives of/for international cooperation, and their organisational and technological implications,
- Open architectures for LRs.

The Conference targets the integration of different types of LRs (spoken, written and other modalities) and of the respective communities. To this end, LREC encourages submissions covering issues which are common to different types of Language Technologies, such as dialog strategy, written and spoken translation, domain-specific data, multimodal communication or multimedia document processing, and will organise, in addition to the usual tracks, common sessions encompassing the different areas of LRs.

Important Dates

Submission of proposals for panels and workshops
20th October 2003

Submission of proposals for oral and poster papers, referenced demos
31st October 2003

Notification of acceptance of workshop and panel proposals
14th November 2003

Notification of acceptance of oral papers, posters, referenced demos
23rd January 2004

Final versions for the proceedings
1st March 2004

NEW RESOURCES

ELRA-W0015 Text Corpus of Le Monde
- All years available now -

All years from Le Monde Text Corpus, starting from 1987, are now available. The prices per year of data are given below:

Price for ELRA members per year of data (research use only)	240.91 Euro
Price for non members per year of data (research use only)	313.18 Euro

ELRA-S0145 Mandarin 5000 database

The MANDARIN-5000 database contains the recordings of 4,752 speakers (2383 males, 2369 females) of Mandarin as first or second language (3,222 native speakers) recorded over the fixed and mobile telephone networks in all provinces of mainland China, including Hong Kong (fixed network: cordless handset: 513 speakers, POT (plain old telephone): 3,558 speakers; mobile network: 491 speakers; undetermined (cordless or mobile): 190 speakers). The database design closely follows the SpeechDat(II) conventions, in particular with respect to the content of the database. The database consists of 1 CD containing all documentation files including the phonetic lexicon, and 3 DVD-R containing the data, i.e. speech files and corresponding transcription files.

Speech samples are stored as sequences of 8-bit 8 kHz A-law, uncompressed. Each prompted utterance is stored in a separate file, and each signal file is accompanied by a transcription file encoded in GB-2312 and ASCII which contains the orthographic representation (i.e. pictograms), phonemic transcription in Pinyin with tones and word boundaries.

Each speaker uttered the following items: 6 isolated application words (25 fixed, 5 free), 1 additional application command with a parameter (e.g. name dialling), 1 sequence of 10 isolated digits (balanced), 6 digit strings (in total balanced for digits, letters, dashes and their transitions), 3 dates, including 1spontaneous one, 2 word spotting phrases using an application word, 2 handset information ("mobile phone ?" "cordless phone ?"), 2 isolated digits, 2 spelled words (letter sequences), 1 currency money amount, 1 natural plain number (balanced for words and transitions), 1 natural number with measure word, 8 names (persons, spelling, cities, companies), where 3 of them spontaneous, 1 spontaneous train schedule request (origin, destination, date, time), 1 spontaneous correction, 1 spontaneous answer to question for time, 1 spontaneous answer to question for time or day, 4 spontaneous answers to questions, including "fuzzy" yes/no, 8 phonetically rich sentences (read newspaper text) for training, and alternatively for test 8 sentences dictated out of newspaper article, 1 time of day (spontaneous), 1 time phrase (read).

The following age distribution has been obtained: 239 speakers are under 16, 2,391 are between 16 and 30, 1,449 are between 31 and 45, 601 are between 46 and 60, and 32 speakers are over 60 (the age of 40 speakers was not determined).

A pronunciation lexicon with orthographic representation (i.e. pictograms), phonemic transcription in Pinyin with tones and frequency of occurrences is also included.

The data collection and transcription were performed by Shanghai Jiao-Tong University in 1999-2000. The owner of the database is Siemens AG, Corporate Technology (Munich, Germany).

	ELRA members	Non-members
For research use	56,000 Euro	70,000 Euro
For commercial use	70,000 Euro	91,500 Euro

ELRA-S0146 Greek SpeechDat-Car database

The Greek SpeechDat-Car database comprises 300 Greek speakers (150 males, 150 females) recorded over the GSM telephone network and in a car. The Greek SpeechDat-Car database was collected and annotated by the Wire Communications Laboratory, Department of Electrical and Computer Engineering of the University of Patras and Knowledge S.A. This database is partitioned into 11 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 boot of the car. The speech data are stored as sequences of 16 kHz, 16 bit and uncompressed. The fifth microphone was connected to the GSM phone, and was recorded on a remote machine, with compressed data stored as sequences of 8 bit A-law 8.kHz. 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 of 5+ digits, 1 spontaneous telephone number, 3 read telephone numbers, 1 credit card number of 14-16 digits, 1 PIN code of 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 embedded application word, 1 question as an extra item, 4 isolated digits, 7 spelled words (1 spontaneous e.g. 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 e.g. own forename or surname), 1 -spontaneous- city of birth/growing up, 2 most frequent cities, 2 most frequent company/agency, 1 "forename surname"), 1 "yes" question, 1 "no" question, 9 phonetically rich sentences, 2 time phrases (1 -spontaneous- time of day, 1-word style- time phrase), 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, and 10 prompts for spontaneous speech.

The following age distribution has been obtained: 185 speakers are between 16 and 30, 79 speakers are between 31 and 45, and 36 speakers are between 46 and 60.

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

	ELRA members	Non-members
For research use	90,000 Euro	120,000 Euro
For commercial use	90,000 Euro	120,000 Euro

ELRA-S0147 Italian Speech Corpus 1 (Appen)

The Italian Speech Corpus 1 contains the recordings of 202 native Italian speakers (112 males, 90 females) recorded in an office and a closed public place, over 4 channels, in a range of low to medium background noise environments (Plantronics Audio 10 (computer/desk mic), Shure SM58 (desk mounted dynamic mic), Shure Beta 53 (headset mic) and Andrea DA-400 (array mic)). The data collection and transcription were performed by Appen (Australia).

Speech samples are stored as sequences of 16-bit 22.05 kHz PCM in uncompressed WAV files.

Each speaker read the following items (prompted): 100 command words, 100 phonetically rich sentences.

The following age distribution has been obtained: 22 speakers are between 18 and 19, 141 are between 20 and 30, 34 are between 31 and 45, and 5 are between 45 and 60. Information about the speakers' place of birth is included.

The database is provided with orthographic transcriptions in SAMPA, including canonical and alternative pronunciation, and syllable, stress and acoustic events markings. All transcriptions were segmented at the utterance (sentence/command word) level, annotated at the word level and checked manually. A pronunciation lexicon including 7,300 headwords (plus variants) is also available. This database is aimed to be used within speech recognition and voice control applications.

	ELRA members	Non-members
For research use	1,200 Euro	1,500 Euro
For commercial use	9,500 Euro	15,000 Euro

ELRA-S0148 Italian TTS Speech Corpus (Appen)

The Italian TTS Speech Corpus contains the recordings of 1 native Italian speaker (male, 50 years old) recorded in a studio over 1 channel (Shure SM15 unidirectional professional head-word condenser microphone). The data collection and transcription were performed by Appen (Australia).

Speech samples are stored as sequences of 16-bit 22.05 kHz PCM in uncompressed WAV files.

The speaker read 3,300 prompted sentences covering all legal triphones and diphones.

The database is provided with orthographic transcriptions in SAMPA, including canonical and alternative pronunciation, and syllable, stress and acoustic events markings. All transcriptions were segmented at the utterance (sentence/command word) level, annotated at the word level and checked manually. A pronunciation lexicon including 7,300 headwords (plus variants) is also available.

This database is aimed to be used within text-to-speech and speech synthesis applications.

	ELRA members	Non-members
For research use	2,000 Euro	3,500 Euro
For commercial use	9,000 Euro	11,000 Euro

ELRA-S0149 Spanish Speech Corpus 1 (Appen)

The Spanish Speech Corpus 1 contains the recordings of 200 native Spanish speakers (100 males, 100 females) recorded in an office and a closed public place, over 4 channels, in a range of low to medium background noise environments (Plantronics Audio 10 (computer/desk mic), Shure SM58 (desk mounted dynamic mic), Shure Beta 53 (headset mic) and Andrea DA-400 (array mic)). The data collection and transcription were performed by Appen (Australia).

Speech samples are stored as sequences of 16-bit 22.05 kHz PCM in uncompressed WAV files.

Each speaker read the following items (prompted): 100 command words, 100 phonetically rich sentences.

The following age distribution has been obtained: 75 speakers are between 18 and 19, 114 are between 20 and 30, and 11 are between 31 and 45. Information about the speakers' place of birth is included.

The database is provided with orthographic transcriptions in SAMPA, including canonical and alternative pronunciation, and syllable, stress and acoustic events markings. All transcriptions were segmented at the utterance (sentence/command word) level, annotated at the word level and checked manually. A pronunciation lexicon including 3,748 headwords (plus variants) is also available. This database is aimed to be used within speech recognition and voice control applications.

	ELRA members	Non-members
For research use	1,200 Euro	1,500 Euro
For commercial use	9,500 Euro	15,000 Euro

ELRA-S0150 Spanish TTS Speech Corpus (Appen)

The Spanish TTS Speech Corpus contains the recordings of 1 native Spanish speaker (male, 28 years old) recorded in a studio over 1 channel (Shure SM15 unidirectional professional head-word condenser microphone). The data collection and transcription were performed by Appen (Australia).

Speech samples are stored as sequences of 16-bit 22.05 kHz PCM in uncompressed WAV files.

The speaker read 1,787 prompted sentences covering all legal triphones and diphones.

The database is provided with orthographic transcriptions in SAMPA, including canonical and alternative pronunciation, and syllable, stress and acoustic events markings. All transcriptions were segmented at the utterance (sentence/command word) level, annotated at the word level and checked manually. A pronunciation lexicon including 3,748 headwords (plus variants) is also available.

This database is aimed to be used within text-to-speech and speech synthesis applications.

	ELRA members	Non-members
For research use	1,500 Euro	1,500 Euro
For commercial use	4,500 Euro	5,500 Euro