The ELRA Newsletter

January - March 2006

Vol.11 n.1

Contents

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

Dear Colleagues,

At the beginning of 2006, ELRA and ELDA have been strongly involved in the preparation of the 5th edition of the Language Resources and Evaluation Conference. LREC 2006 took place in Genoa, from 22nd to 29th May. The next ELRA newsletter, which should be distributed shortly, will give you an overview of the conference, with some sessions' and workshops' summaries.

Early May, the ELRA Annual General Assembly was held in Paris. During this meeting, chaired by Bente Maegaard, President of ELRA and attended by 13 members, ELRA activities were reviewed and financial data (2005 report and 2006 budget) presented. In addition, as stated in the ELRA statutes, at the end of each term of two years, the ELRA Board was renewed, with 5 new members: representatives from Polderland Language & Speech Technology (Netherlands), Morphologic (Hungary), ILSP (Greece), Nuance (Belgium), DFKI (Germany). We would like to thank the Board members who left for their contribution to the success and advances at ELRA.

As for this newsletter, in "Announcing the AMI Corpus", it presents the AMI Meeting Corpus that has been collected in the framework of the European-funded AMI project (FP6-506811) "dedicated to the research and development of technology that will augment communications between individuals and groups of people". It also contains an overview of the European project CHIL - Computers in the Human Interaction Loop.

New resources have been secured for distribution. These are announced in the last section of this newsletter and consist of :

Speech Telephone
- S0191: ZipTel

Speech Desktop/Microphone
- S0192: GlobalPhone Arabic
- S0193: GlobalPhone Chinese-Mandarin
- S0194: GlobalPhone Chinese-Shanghai
- S0195: GlobalPhone Croatian
- S0196: GlobalPhone Czech
- S0197: GlobalPhone French
- S0198: GlobalPhone German
- S0199: GlobalPhone JaPanese
- S0200: GlobalPhone Korean
- S0201: GlobalPhone Portuguese (Brazilian)
- S0202: GlobalPhone Russian
- S0203: GlobalPhone Spanish (Latin American)
- S0204: GlobalPhone Swedish
- S0205: GlobalPhone Tamil
- S0206: GlobalPhone Turkish
- S0209: Oxford English phonetics files
- S0210: Shorter Oxford English Dictionary - Audio Files
- S0211: USpanish Speecon Database
- S0212: Taiwain Mandarin Speecon database
- S0213: Italian Speecon Database
- S0214: Swedish Speecon Database

Speech related
- S0207: LC-STAR Catalan phonetic lexicon of proper names
- S0208: LC-STAR Spanish phonetic lexicon of proper names

Written Corpora
- W0040: Venice Italian Treebank (VIT)
- W0041: Corpus of Contemporaneous Spanish Novels

Monolingual Lexica
- L0058: British English Source Lexicon (BESL) version 2.2
- L0059: Offensive Word Filter 1
- L0060: Offensive Word Filter 2
- L0061: The Oxford Spanish Dictionary
- L0062: French Source Lexicon

Terminology Resources
- T0368: Multilingual Wordbank
- T0369: Multilingual Phrasebank
- T0370: Dictionary of Law
- T0371: Dictionary of Medicine

Evaluation Packages
- E0002: TC-STAR Evaluation Package ASR English
- E0003: TC-STAR Evaluation Package ASR Spanish
- E0004: TC-STAR Evaluation Package ASR Mandarin Chinese
- E0005: TC-STAR Evaluation Package SLT English to Spanish
- E0006: TC-STAR Evaluation Package SLT Spanish to English
- E0007: TC-STAR Evaluation Package SLT Chinese to English

Once again, if you would like to join ELRA and benefit from its services (that are summarized at www.elra.info), please, do not hesitate to contact us.

Bente Maegaard, President

Khalid Choukri, CEO

Announcing the AMI Meeting Corpus

Jean Carletta

The European-funded AMI project (FP6-506811) is a 15-member multi-disciplinary consortium dedicated to the research and development of technology that will help groups interact better. One AMI focus is on developing meeting browsers that improve work group effectiveness by giving better access to the group's history. Increasingly in future, we will be considering how related technologies can help group members joining a meeting late or having to "attend" from a different location. In both cases, a key part of our approach is to index meetings for the properties that users find salient. This might mean, for instance, spotting topic boundaries, decisions, intense discussions, or places where a specific person or subject was mentioned. To help with developing this indexing the consortium has collected the AMI Meeting Corpus, a set of recorded meetings that is now available as a public resource. Although the data set was designed specifically for the project, it could be used for many different purposes in linguistics, organizational and social psychology, speech and language engineering, video processing, and multi-modal systems.

The AMI Meeting Corpus consists of 100 hours of meeting recordings. The recordings use a range of signals synchronized to a common timeline. These include close-talking and far-field microphones, individual and room-view video cameras, and output from a slide projector and an electronic whiteboard. During the meetings, the participants also have unsynchronized pens available to them that record what is written. The meetings were recorded in English using three different rooms with different acoustic properties, and include mostly non-native speakers. Figure 1 illustrates one of these rooms.

The most useful speech corpora are those that come with annotations. The AMI Meeting Corpus includes high quality, manually produced orthographic transcription for each individual speaker, including word-level timings that have derived by using a speech recognizer in forced alignment mode. It also contains a wide range of other annotations, not just for linguistic phenomena but also detailing behaviours in other modalities. These include dialogue acts; topic segmentation; extractive and abstractive summaries; named entities (refer to Figure 2 for a sample use of the annotation tool); the types of head gesture, hand gesture, and gaze direction that are most related to communicative intention; movement around the room; emotional state; and where heads are located on the video frames. The linguistically motivated annotations have been applied the most widely, and cover all of the scenario-based recordings. Other annotations are more limited, but in each case we have chosen what we consider a sensible data subset. For phenomena that are sparse in the meeting recordings, we have marked up auxiliary recordings where the behaviours are more common. These are also included in the corpus.

Figure 1: One of AMI's three instrumented meeting rooms.

There are a number of features that make the AMI Meeting Corpus a bit different from previous corpora. The first is its release under a Creative Commons Attribution ShareAlike Licence. This form of licensing is intended to create an environment in which people freely share what they have created. The corpus license allows users to copy, distribute, and display the data for non-commercial purposes as long as the AMI project is credited. However, if the user wishes to distribute anything derived using the corpus, that can only be done under the same license as the original data. Although Creative Commons licensing is relatively new for data sets, it is similar to the GNU General Public License, which is already in common use for research software. The license does not bar us from distributing the data under other terms as well, but it does allow us to give the data away to the widest group possible without fear of being exploited. In June, we released the entire database of signals under a ShareAlike license, along with the orthographic transcription and some of the annotations. The remaining annotations are scheduled for release in stages, with the complete set scheduled for January 2007.

Another feature of note is that unlike for most previous corpora, all of the annotations provided are in one consistent format that represents not just the time course of the annotations, but also how they relate structurally to the transcription and to other annotations. For instance, topic segments and dialogue acts are represented not just as labelled spans with a start and end time, but as timed sequences of words. Extractive summaries do not just pull out segments of the meetings by time, but point to the dialogue act (and from there, the words) to be extracted, as well as any sentences in the meeting abstract that relate to the extracted segment. This kind of representation, which can be built and searched using the open source NITE XML Toolkit (http://www.ltg.ed.ac.uk/NITE), allows for much richer investigation of the data than is possible using simple time stamping. It also makes data sharing a more attractive proposition, since it makes integrating new annotations easier and increases their possible uses. In the past, annotations have been created by many different sites for popular data sets like the Switchboard Corpus, but sharing has been patchy. Our license terms and data representation are both designed to move the community to a model in which we pool our resources better.

Finally, the AMI Meeting Corpus has a somewhat unusual design. Except for corpora set up to inform a spoken dialogue systems application by showing what the system needs to produce, dialogue corpus designers usually aim to capture completely natural, uncontrolled conversations. Around one-third of our data is like this; it consists of meetings from various groups that would have happened whether they were being recorded or not. However, the rest has been collected by having the participants play different roles in a fictitious design team that takes a new project from kick-off to completion over the course of a day. The day starts with training for the participants about what is involved in the roles they have been assigned (industrial designer, interface designer, marketing, or project manager) and then contains four meetings, plus individual work to prepare for them and to report on what happened. All of their work is embedded in a very mundane work environment that includes web pages, email, text processing, and slide presentations.

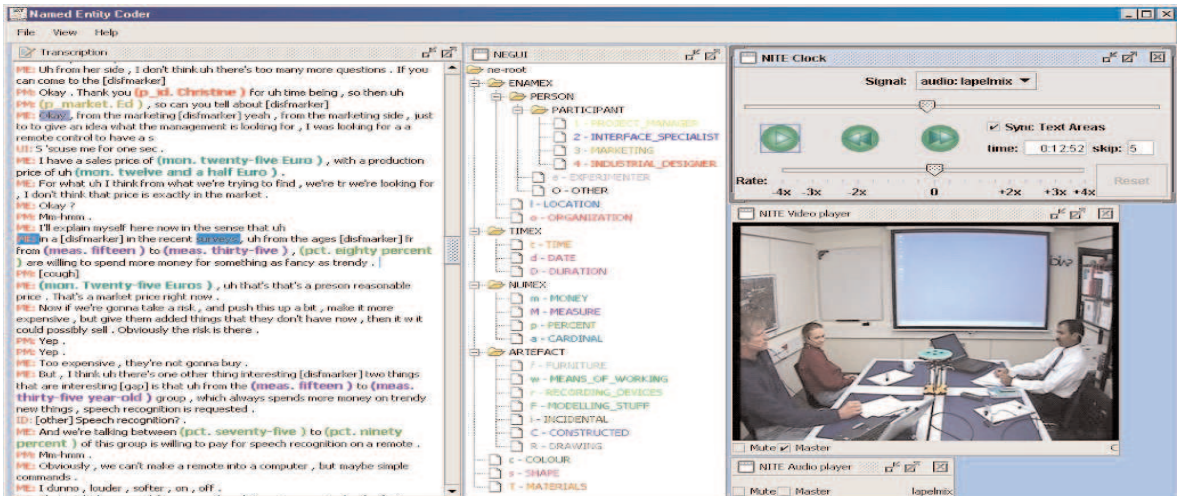

Figure 2: The NXT-based coding tool used to create named entity annotation for the AMI Meeting Corpus.

This sort of role-playing has some striking advantages over natural data, particularly for research that relies on the meaning of what was said. First and foremost, measures for the quality of a group outcome can be built in. Valid measures are very difficult to obtain for natural groups, but they are invaluable for assessing whether technologies for assisting human groups actually help. Although none of the groups in the AMI Meeting Corpus use browsers to access their past meetings, we intend to collect groups that do in the near future so that we can compare their outcomes both to unassisted groups and to each other. In addition, the behaviour of role-playing groups is easier to understand than that of natural groups. This is because the researcher is in control of the participants' knowledge and motivation, and because the groups don't come with years of personal history that cause them to behave differently than they would otherwise. The usual disadvantage expressed for role-playing is that there is no guarantee participants will care enough about what they do to provide data comparable to natural interaction. In our experience, as long as the role-play is set up carefully, participants in role-playing fully engage in what they are doing. However, the reason why the corpus is not completely made up of controlled data is as a safeguard, both against any possible disadvantages of role-playing and against the domain limitations it entails. The inclusion of both controlled and natural data allows researchers to develop new techniques on the controlled data first and then begin to test their generalizability using the natural material.

Our main way of releasing the corpus is through the website (http://corpus.amiproject.org) see Figure 3. At the website, anyone can look at the signals for one meeting and read extensive documenta-

tion. After registration, users can browse meetings online using SMIL presentations, download their chosen data, and participate in a discussion forum. Registration is simple and free. Everything that has been released is on the website, apart from the full-size videos. These are too large for download, but the website gives a contact for ordering firewire drives that contain them, priced at the cost of production. In addition to the website, we have also produced 500 copies of a "taster" DVD that includes everything for a single meeting - signals, transcripts, and every available annotation, including samples of some types that have not yet formed part of the public release. The DVD can currently be ordered for free from the website.

The AMI Meeting Corpus has required substantial investment. We expect it to become an invaluable resource to the broader research community, as it provides novel data for research and evaluation in many different areas. The AMI project consortium will continue working together in the newly-funded AMIDA project, and therefore intends both to maintain the corpus and to take an interest in its growth. We are happy to have it used, and hope that it will attract researchers with other approaches to our own problems, but also be taken in new and unforeseen directions.

Jean Carletta
Senior Research Fellow at the Institute for Communicating and Collaborative Systems,
University of Edinburgh, UK.
jeanc@inf.ed.ac.uk

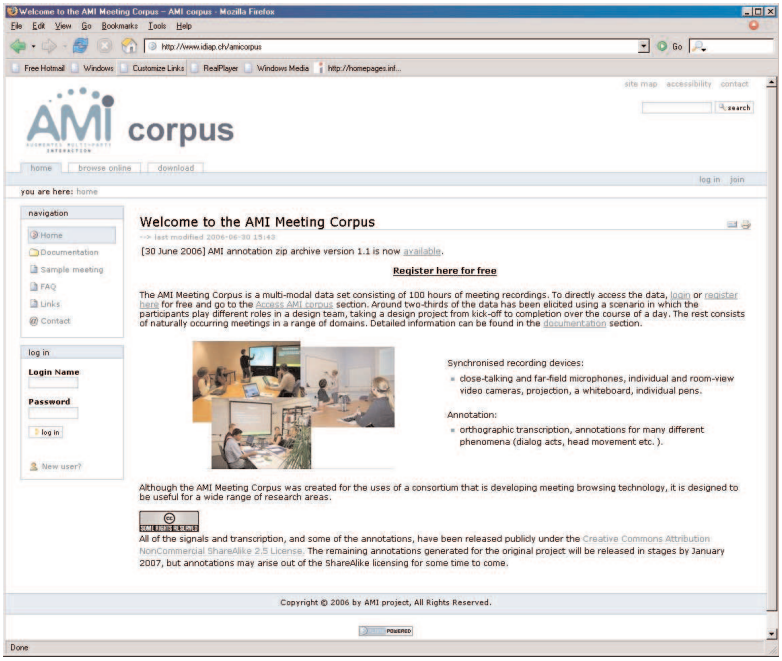

Figure 3: A screenshot of the AMI corpus website.

CHIL - Computers in the Human Interaction Loop - Electronic Butlers simplify your daily life

Margit Rödder

A cell phone ringing in the theatre, trying for hours to reach someone on the phone, attending a meeting and forgetting the documents, endless discussions, forgetting the name of a partner or friend… - that will soon be history. CHIL provides useful proactive and intelligent services, which will cause a fundamental shift in the way we use computers today. We aim to realize computer services that are delivered to people in an implicit, indirect and unobtrusive way. Computers in the Human Interaction Loop (CHIL) aims to introduce computers into a loop of humans interacting with humans, rather than condemning a human to operate in a loop of computers. This will give humans the most valuable gift: more time.

CHIL Scenario

A CHIL scenario is a situation in which people interact face to face with people, exchange information, collaborate to jointly solve problems, learn, or socialize, using all their natural ways of face to face communication (speech/language, gestures, body posture, etc.). Therefore the focus in the project is on two scenarios: offices and lecture rooms.

CHIL Services

In order to provide really useful proactive and intelligent services, the Who, Where, What, Why and How of Human activities and communication needs to be perceived and understood. This presents a fundamental departure from the way we have thought of user interfaces up till now. It requires robust, multimodal perceptual interfaces capable of tracking, identifying, recognizing and understanding the role, purpose and content of human communication, activities, state and their environment. If these machines in human contexts were available, a new class of digital services could be developed that would take concrete advantage of these new capabilities. CHIL explores four particular services as instantiations of this vision:

Memory Jog

The Memory Jog provides attendees with information related to a situation (e.g. a meeting or a lecture) and related to the participants in it. It provides context- and content-aware information pull and push, both personalized and public.

Relational Report

This report evaluates the individual's contribution to the group's activity. Multimedia reports about the relational behaviour of each participant are privately delivered as part of an automatic coaching system. The idea is that the whole organization might benefit from an increase of awareness of participants about their own behaviours during group activities.

Figure 1: The Collaborative worksspace supports synchronous cooperation and participation by providing a shared information space and tools for managing face-to-face meeting.

Connector

The Connector is a context-aware connecting service ensuring that two parties get connected by the most appropriate media at the right time and place. Based on the observed context, and each party's preference, it decides when and how it is most appropriate and desirable for both parties to be connected.

Socially-Supportive Workspaces

Socially-Supportive Workspaces are an infrastructure for fostering cooperation among participants, whereby the system provides a multimodal interface for entering and manipulating contributions from different participants, e.g., enabling joint discussion of minutes, or joint accomplishment of a common task, with people proposing their ideas, and making them available on the shared workspace, where they are discussed by the whole group. The Socially-Supportive Workspaces provide a facilitator functionality that is able to monitor group activities to keep it on track, such as suggesting moving on to

Figure 2: The Connector Devices bring two parties together at the most appropriate time.

the next task, and can better support social relationships.

CHIL Technologies

In order to develop the described services, it is necessary to continuously track human activities, using all perception modalities available, and build static and dynamic models of the scene. With the different technologies, user profiles must be learned and behavioural patterns detected. The perceived multimodal information must be combined to better analyze the scene and to provide pertinent assistance.

In pursuing its target, the CHIL technique develops innovations advancing the state of the art in a wide range of component technologies:

- Audio Visual Person Tracking
- Pointing Gesture Recognition
- Face Detection and Face Recognition
- Head Pose Estimation
- Collaborative Workspace for Meetings
- Automatic Meeting Summarizer
- Agent Based Software Architecture
- Context Aware Management of Communication
- Animated Secretary - improved communication using virtual talking heads
- "SitCom" - the tool for situation modelling and visualization
- "AVASR" - the technology prototype for audio visual automatic speech recognition using far-field or close-talking audio-visual sensors
- "targeted audio", an array of small ultrasound speakers, that can deliver a very focussed audio beam.

CHIL - Software Architecture

Realizing the goal of the project demands that perceptual interfaces are integrated according to the design, purpose and objectives of the targeted services. Rather than focusing on an ad-hoc implementation of particular services, the CHIL project proceeds by specifying a structured method for interfacing with sensors, integrating technology components, processing sensorial input and ultimately composing non-obtrusive services as collections of basic service capabilities. Moreover, it enables management of multimodal user interactions. Thus, in terms of software infrastructure the architecture supports components communication and multimodal interactions. Based on this infrastructure, strategies for situation detection, assessment and decision-making are implemented.

Outlook

CHIL shows a vision of the future - a new approach to more supportive and less burdensome computing and communication services. The international and multidisciplinary team sets out to study the technical, social and ethical questions that will enable this next generation of computing in a responsible manner.

CHIL-PARTNERS

Fifteen partners from nine countries in Europe and the US collaborate in the CHIL consortium:

- Fraunhofer-Institut für Informations- und Datenverarbeitung (IITB), Germany
- Universität Karlsruhe (TH), Interactive Systems Labs (ISL), Germany
- DaimlerChrysler AG, Group Dialogue Systems, Germany
- Evaluations and Language resources Distribution Agency (ELDA), France
- IBM Ceska Republika, Czech Republic
- Research and Education Society in Information Technologies (RESIT), Greece
- Institut National de Recherche en Informatique et en Automatique (INRIA), Lab GRAVIR, France
- Istituto Trentino di Cultura (IRST), Italy
- Kungl Tekniska Högskolan (KTH), Sweden

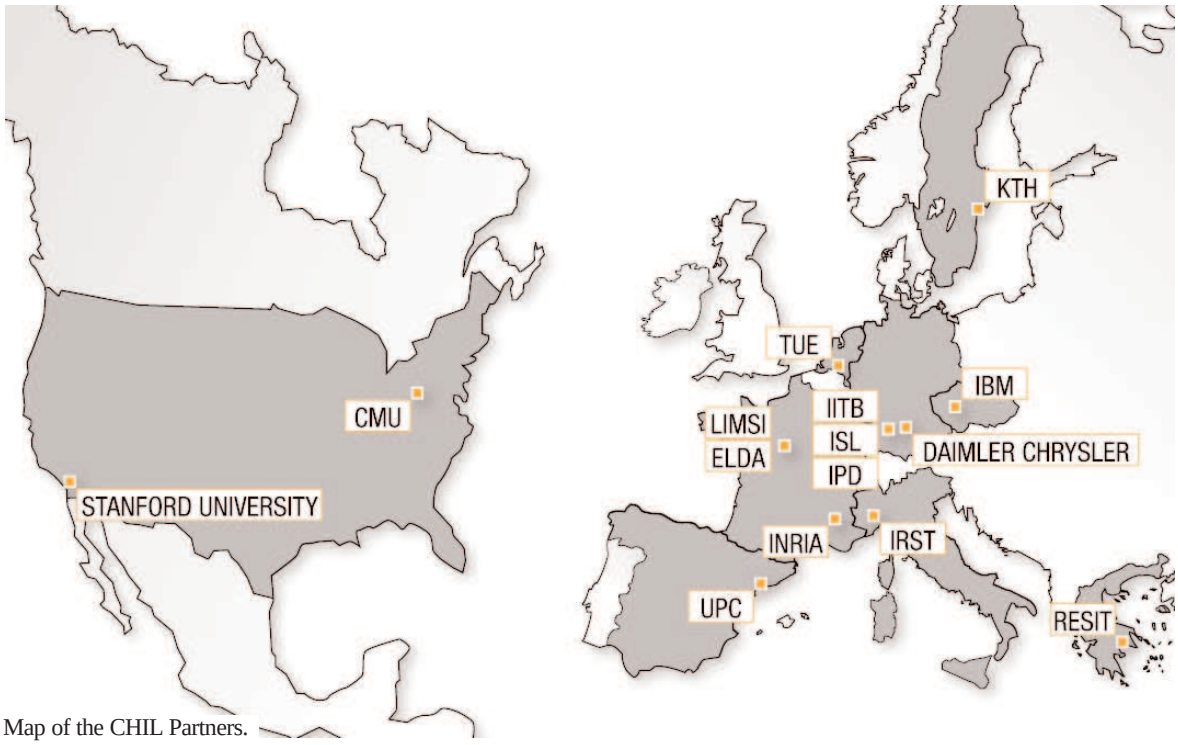

Figure 3: Map of the CHIL Partners.

- Centre National de la Recherche Scientifique, (CNRS), LIMSI, France
- Technische Universiteit Eindhoven (TUE), The Netherlands
- Universität Karlsruhe (TH), Institute for Program Structures and Data Organisation, (IPD), Germany
- Universitat Politècnica de Catalunya (UPC), Spain
- Stanford University, USA
- Carnegie Mellon University (CMU), USA

This Integrated Project CHIL IP 506909 is supported by funding in the thematic area Information Society Technologies under the Sixth Research Framework Programme of the European Union.

Further information under:

http://chil.server.de

Scientific Coordinator
Universität Karlsruhe (TH)
Interactive Systems Labs
http://isl.ira.uka.de
Prof. Alex Waibel, ahw@cs.cmu.edu
Dr. Rainer Stiefelhagen, stiefel@ira.uka.de

New Resources

ELRA-S0191 ZipTel

The ZipTel telephone speech database contains recordings of people applying for a SpeechDat prompt sheet via telephone. For the SpeechDat data collection, calls for participation were published in "phone", the customer magazine of the mobile telephone provider "e-plus", and in numerous newspapers all over Germany. In these calls, a telephone number was given where callers could order a SpeechDat prompt sheet. The calls were recorded by an automatic telephone server; callers were asked to provide address, ZIP code, city and telephone number.

Total number of recordings: 7746

Total duration: 14h

Format: SpeechDat Exchange Format, SAM, BAS Partitur Format (BPF)

	ELRA members	Non-members
For research use	627.17 Euro	754.35 Euro
For commercial use	4,627.17 Euro	4,754.35 Euro

GlobalPhone databases

GlobalPhone is a multilingual speech and text database collected at Karlsruhe University, Germany. The GlobalPhone corpus provides transcribed speech data for the development and evaluation of large vocabulary continuous speech recognition systems in the most widespread languages of the world. GlobalPhone is designed to be uniform across languages with respect to the amount of text and audio per language, the audio data quality (microphone, noise, channel), the collection scenario (task, setup, speaking style etc.), and the transcription conventions. As a consequence, GlobalPhone supplies an excellent basis for research in the areas of (1) multilingual speech recognition, (2) rapid deployment of speech processing systems to new languages, (3) language and speaker identification tasks, as well as (4) monolingual speech recognition in a large variety of languages.

To date, the GlobalPhone corpus covers 15 languages Arabic (Modern Standard Arabic), Chinese-Mandarin, Chinese-Shanghai, Croatian, Czech, French, German, Japanese, Korean, Portuguese (Brazilian), Russian, Spanish (Latin American), Swedish, Tamil, and Turkish. This selection covers a broad variety of language peculiarities relevant for Speech and Language Research and Development. It comprises widespread languages (Arabic, Chinese, Spanish), contains economically and politically important languages (Korean, Japanese, Arabic), and spans over wide geographical areas (Europe, America, Asia). The spoken speech covers a wide selection of phonetic characteristics, e.g. tonal sounds (Mandarin, Shanghai), pharyngal sounds (Arabic), consonantal clusters (German), nasals (French, Portuguese), palatized sounds (Russian), and more. The written language contains large orthographic variations, such as phonologic scripts (alphabetic scripts such as Roman, Cyrillic, Arabic; syllable-based scripts like Japanese Kana, Korean Hangul), and ideographic scripts (Chinese Hanzi and Japanese Kanji). The languages cover many morphological variations, e.g. agglutinative languages (Turkish, Korean), compounding languages (German), and also include scripts that completely lack word segmentation (Chinese).

The data acquisition was performed in countries where the language is officially spoken. In each language about 100 adult native speakers were asked to read 100 sentences. The read texts were selected from national newspaper articles available from the web to cover a wide domain with large vocabulary. The articles report national and international political news, as well as economic news mostly from the years 1995-1998. The speech data was recorded with a Sennheiser 440-6 close-speaking microphone and is available in identical characteristics for all languages: PCM encoding, mono quality, 16-bit quantization, and 16 kHz sampling rate. Most of the speech data was recorded in a quiet office, some are recorded in apartments, i.e. living room. The transcriptions are available in the original script of the corresponding language. In addition, all transcriptions have been romanized, i.e. transformed into Roman script applying customized mapping algorithms. The transcripts are validated and supplemented by special markers for spontaneous effects like stuttering, false starts, and non-verbal effects such as breathing, laughing, and hesitations. Speaker information, such as age, gender, occupation, etc. as well as information about the recording

setup complement the database. The entire GlobalPhone corpus contains over 300 hours of speech spoken by more than 1500 native adult speakers. The data are divided in speaker disjoint sets for training, development, and evaluation (80:10:10) and are organized by languages and speakers.

The list of available GlobalPhone resources is given below:

ELRA-S0192 GlobalPhone Arabic

The Arabic corpus was produced using the Assabah newspaper. It contains recordings of 78 speakers (35 males, 43 females) recorded in Tunisia, Palestine and Jordan. The following age distribution has been obtained: 20 speakers are below 19, 35 speakers are between 20 and 29, 13 speakers are between 30 and 39, 6 speakers are between 40 and 49, and 4 speakers are over 50.

ELRA-S0193 GlobalPhone Chinese-Mandarin

The Chinese-Mandarin corpus was produced using the Peoples Daily newspaper. It contains recordings of 132 speakers (64 males, 68 females) recorded in Beijing, Wuhan and Hekou, China. The following age distribution has been obtained: 16 speakers are below 19, 96 speakers are between 20 and 29, 16 speakers are between 30 and 39, 3 speakers are between 40 and 49 (1 speaker age is unknown).

ELRA-S0194 GlobalPhone Chinese-Shanghai

The Chinese-Shanghai corpus was produced using the Peoples Daily newspaper. It contains recordings of 41 speakers (16 males, 25 females) recorded in Shanghai, China. The following age distribution has been obtained: 1 speaker is below 19, 2 speakers are between 20 and 29, 13 speakers are between 30 and 39, 14 speakers are between 40 and 49, and 11 speakers are over 50.

ELRA-S0195 GlobalPhone Croatian

The Croatian corpus was produced using the HRT and Obzor Nacional newspapers. It contains recordings of 94 speakers (38 males, 56 females) recorded in Zagreb, Croatia, and parts of Bosnia. The following age distribution has been obtained: 21 speakers are below 19, 30 speakers are between 20 and 29, 14 speakers are between 30 and 39, 15 speakers are between 40 and 49, and 13 speakers are over 50 (1 speaker age is unknown).

ELRA-S0196 GlobalPhone Czech

The Czech corpus was produced using the Ceskomoravsky Profit Journal and Lidove Noviny newspaper. It contains recordings of 102 speakers (57 males, 45 females) recorded in Prague, Czech Republic. The following age distribution has been obtained: 16 speakers are below 19, 70 speakers are between 20 and 29, 2 speakers are between 30 and 39, 9 speakers are between 40 and 49, and 5 speakers are over 50.

ELRA-S01967 GlobalPhone French

The French corpus was produced using Le Monde newspaper. It contains recordings of 100 speakers (49 males, 51 females) recorded in Grenoble, France. The following age distribution has been obtained: 3 speakers are below 19, 52 speakers are between 20 and 29, 16 speakers are between 30 and 39, 13 speakers are between 40 and 49, and 14 speakers are over 50 (2 speakers age is unknown).

ELRA-S0198 GlobalPhone German

The German corpus was produced using the Frankfurter Allgemeine und Sueddeutsche Zeitung newspaper. It contains recordings of 77 speakers (70 males, 7 females) recorded in Karlsruhe, Germany. No age distribution is available.

ELRA-S0199 GlobalPhone Japanese

The Japanese corpus was produced using the Nikkei Shinbun newspaper. It contains recordings of 149 speakers (104 males, 44 females, 1 unspecified) recorded in Tokyo, Japan. The following age distribution has been obtained: 22 speakers are below 19, 90 speakers are between 20 and 29, 5 speakers are between 30 and 39, 2 speakers are between 40 and 49, and 28 speakers are over 50 (2 speakers age is unknown).

ELRA-S0200 GlobalPhone Korean

The Korean corpus was produced using the Hankyoreh Daily News. It contains recordings of 100 speakers (50 males, 50 females) recorded in Seoul, Korea. The following age distribution has been obtained: 7 speakers are below 19, 70 speakers are between 20 and 29, 19 speakers are between 30 and 39, and 3 speakers are between 40 and 49 (1 speaker age is unknown).

ELRA-S0201 GlobalPhone Portuguese (Brazilian)

The Portuguese (Brazilian) corpus was produced using the Folha de Sao Paulo newspaper. It contains recordings of 102 speakers (54 males, 48 females) recorded in Porto Velho and Sao Paulo, Brazil. The following age distribution has been obtained: 6 speakers are below 19, 58 speakers are between 20 and 29, 27 speakers are between 30 and 39, 5 speakers are between 40 and 49, and 5 speakers are over 50 (1 speaker age is unknown).

ELRA-S0202 GlobalPhone Russian

The Russian corpus was produced using the Ogonyok Gaseta and Express-Chronika newspapers. It contains recordings of 115 speakers (61 males, 54 females) recorded in Minsk, Belarus. The following age distribution has been obtained: 9 speakers are below 19, 76 speakers are between 20 and 29, 9 speakers are between 30 and 39, 15 speakers are between 40 and 49, and 6 speakers are over 50.

ELRA-S0203 GlobalPhone Spanish (Latin American)

The Spanish (Latin America) corpus was produced using the La Nacion newspaper. It contains recordings of 100 speakers (44 males, 56 females) recorded in Heredia and San Jose, Costa Rica. The following age distribution has been obtained: 20 speakers are below 19, 54 speakers are between 20 and 29, 13 speakers are between 30 and 39, 5 speakers are between 40 and 49, and 8 speakers are over 50.

ELRA-S0204 GlobalPhone Swedish

The Swedish corpus was produced using the Goeteborgs-Posten newspaper. It contains recordings of 98 speakers (50 males, 48 females) recorded in Stockholm and Vaernamo, Sweden. The following age distribution has been obtained: 9 speakers are below 19, 50 speakers are between 20 and 29, 12 speakers are between 30 and 39, 11 speakers are between 40 and 49, and 16 speakers are over 50.

ELRA-S0205 GlobalPhone Tamil

The Tamil corpus was produced using the Thinaboomi Tamil Daily newspaper. It contains recordings of 47 speakers (gender unspecified) recorded in India. No age distribution is available.

ELRA-S0206 GlobalPhone Turkish

The Turkish corpus was produced using the Zaman newspaper. It contains recordings of 100 speakers (28 males, 72 females) recorded in Istanbul, Turkey. The following age distribution has been obtained: 30 speakers are below 19, 30 speakers are between 20 and 29, 23 speakers are between 30 and 39, 14 speakers are between 40 and 49, and 3 speakers are over 50.

PRICES

• For S0194

	ELRA members	Non-members
For research use	300 Euro	355 Euro
For commercial use	1,800 Euro	2,125 Euro

• For S0192, S0193, S0195, S0196, S0197, S198, S0199, S0200, S0201, S202, S203, S204, S0206

	ELRA members	Non-members
For research use	600 Euro	700 Euro
For commercial use	3,000 Euro	3,600 Euro

• For S0205

	ELRA members	Non-members
For research use	100 Euro	125 Euro
For commercial use	500 Euro	600 Euro

Special prices for a purchase of several GlobalPhone Languages

• 5 Languages

	ELRA members	Non-members
For research use	2,600 Euro	3,000 Euro
For commercial use	13,500 Euro	16,200 Euro

• 10 Languages

	ELRA members	Non-members
For research use	5,000 Euro	6,000 Euro
For commercial use	24,000 Euro	28,800 Euro

• 15 Languages

	ELRA members	Non-members
For research use	7,500 Euro	9,000 Euro
For commercial use	31,500 Euro	37,800 Euro

ELRA-S0207 LC-STAR Catalan phonetic lexicon of proper names

The LC-STAR Catalan phonetic lexicon of proper names was created within the scope of the LC-STAR project (IST 2001-32216) which was sponsored by the European Commission and the Spanish Government.

The lexicon comprises a set of more than 45,000 proper names (including person names, family names, cities, streets, companies and brand names) with phonetic transcriptions in SAMPA.

The lexicon is distributed in one CD-ROM.

	ELRA members	Non-members
For research use	9,250 Euro	14,000 Euro
For commercial use	15,000 Euro	18,750 Euro

ELRA-S0208 LC-STAR Spanish phonetic lexicon of proper names

The LC-STAR Spanish phonetic lexicon of proper names was created within the scope of the LC-STAR project (IST 2001-32216) which was sponsored by the European Commission and the Spanish Government.

The lexicon comprises a set of more than 45,000 proper names (including person names, family names, cities, streets, companies and brand names) with phonetic transcriptions in SAMPA.

The lexicon is distributed in one CD-ROM.

	ELRA members	Non-members
For research use	9,250 Euro	14,000 Euro
For commercial use	15,000 Euro	18,750 Euro

ELRA-S0209 Oxford English phonetics files

Derived from a range of Oxford Dictionaries, these files list word forms together with a representation of their IPA pronunciation. It contains 250,000 words. Pronunciation is based on standard British English. Word forms include dictionary lemmas and inflections or other morphological variations, plus a wide range of proper name and encyclopedic material. The data also includes a large number of common multiword phrases and compound nouns.

The files are provided in XML.

	ELRA members	Non-members
For research use	4,000 Euro	5,000 Euro

ELRA-S0210 Shorter Oxford English Dictionary - Audio Files

These are recorded headwords for the Shorter Oxford English Dictionary. British English pronunciation.

Coverage: over 95,000 soundfiles, average filesize 10KB

Features: high-quality recordings with British English pronunciations, accurate coverage of different homographs, variant forms and inflections, clear linking of soundfiles to phonetic information, full information on parts of speech and subsenses covered

Format: 11kHz 8-bit WAV

	ELRA members	Non-members
For research use	7,000 Euro	10,000 Euro

ELRA-S0211 US Spanish Speecon Database

The US Spanish Speecon database is divided into 2 sets:
1) The first set comprises the recordings of 550 adult Spanish speakers in the US (255 males, 295 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).
2) The second set comprises the recordings of 50 child Spanish speakers in the US (28 boys, 22 girls), recorded over 4 microphone channels in 1 recording environment (children room).
This database is partitioned into 22 DVDs (first set) and 3 DVDs (second set).
The speech databases made within the Speecon project were validated by SPEX, the Netherlands, to assess their compliance with the Speecon format and content specifications.
Each of the four speech channels is recorded at 16 kHz, 16 bit, uncompressed unsigned integers in Intel format (lo-hi byte order).
To each signal file corresponds an ASCII SAM label file which contains the relevant descriptive information.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro

ELRA-S0212 Taiwan Mandarin Speecon database

The Taiwan Mandarin Speecon database is divided into 2 sets:
1) The first set comprises the recordings of 550 adult Taiwanese speakers (273 males, 277 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).
2) The second set comprises the recordings of 50 child Taiwanese speakers (25 boys, 25 girls), recorded over 4 microphone channels in 1 recording environment (children room).
This database is partitioned into 56 DVDs (first set) and 3 DVDs (second set).
The speech databases made within the Speecon project were validated by SPEX, the Netherlands, to assess their compliance with the Speecon format and content specifications.
Each of the four speech channels is recorded at 16 kHz, 16 bit, uncompressed unsigned integers in Intel format (lo-hi byte order).
To each signal file corresponds an ASCII SAM label file which contains the relevant descriptive information.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro

ELRA-S0213 Italian Speecon Database

The Italian Speecon database is divided into 2 sets:
1) The first set comprises the recordings of 550 adult Italian speakers (273 males, 277 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).
2) The second set comprises the recordings of 50 child Italian speakers (28 boys, 22 girls), recorded over 4 microphone channels in 1 recording environment (children room).
This database is partitioned into 23 DVDs (first set) and 3 DVDs (second set).
The speech databases made within the Speecon project were validated by SPEX, the Netherlands, to assess their compliance with the Speecon format and content specifications.
Each of the four speech channels is recorded at 16 kHz, 16 bit, uncompressed unsigned integers in Intel format (lo-hi byte order).
To each signal file corresponds an ASCII SAM label file which contains the relevant descriptive information.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro

ELRA-S0214 Swedish Speecon database

The Swedish Speecon database is divided into 2 sets:
1) The first set comprises the recordings of 550 adult Swedish speakers (270 males, 280 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).
2) The second set comprises the recordings of 50 child Swedish speakers (25 boys, 25 girls), recorded over 4 microphone channels in 1 recording environment (children room).
This database is partitioned into 23 DVDs (first set) and 3 DVDs (second set).
The speech databases made within the Speecon project were validated by SPEX, the Netherlands, to assess their compliance with the Speecon format and content specifications.
Each of the four speech channels is recorded at 16 kHz, 16 bit, uncompressed unsigned integers in Intel format (lo-hi byte order).
To each signal file corresponds an ASCII SAM label file which contains the relevant descriptive information.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro

ELRA-W0040 Venice Italian Treebank (VIT)

The VIT, Venice Italian Treebank is the effort of the collaboration of people working at the Laboratory of Computational Linguistics of the University of Venice in the years 1995-2005. It is partly the result of annotation carried out internally with no specific project in mind and no financial support. This work was partly related to the development of a lexicon, a morphological analyzer, a tagger, a deep parser of Italian. All these resources were finally ready at the beginning of the '90s when the LCL got involved in the first national projects.

The VIT contains about 272,000 words distributed over six different domains, and this is what makes it so relevant for the study of the structure of Italian language. The following domains were annotated: Domain Number of words Time span, Bureaucratic 20,000 1986, Politics 40,000 1984, Economic & financial 12,000 1987, Literary 10,000 1984, Scientific 20,000 1985, News 170,000 1994.

In addition, some 60,000 tokens of spoken dialogues in different Italian varieties were annotated.

The annotation follows general X-bar criteria with 29 constituency labels and 102 PoS tags. VIT is also made available in a broad annotation version with 10 constituency labels and 22 PoS tags for machine learning purposes.

The format is plain text with square bracketing. However, a UPenn style version which is readable by the open source query language CorpusSearch is also provided.

	ELRA members	Non-members
For research use	3,000 Euro	4,000 Euro
For commercial use	7,000 Euro	10,000 Euro

ELRA-W0041 Corpus of Contemporaneous Spanish Novels

This corpus consists of 11 novels written in Castilian Spanish by Inmaculada Ferrer-Vidal Turull, a contemporaneous author. The list of novels consists of: La búsqueda: 113,639 words, Tristeza: 41,125 words, Cuarto menguante: 42,419 words, Recuerdos: 55,694 words, Sucedió en Abril: 46,040 words, Viejos amigos: 84,082 words, Soledad & Cia: 69,848 words, El chispazo, la hoguera y las brasas: 108,877 words, Un giro en la vida : 70,736 words, Adiós: 2,016 words, Vacaciones: 3,623 words

The novels are available in Word format.

	ELRA members	Non-members
For research use	400 Euro	500 Euro
For commercial use	800 Euro	1,000 Euro

ELRA-L0058 British English Source Lexicon (BESL) version 2.2

BESL is a complete database of the English lexicon. It consists of over 230,000 lemmas, over 350,000 word forms, 60,000 proper nouns, 3,000 abbreviations, and 58,000 multi-word compound nouns. Each headword is provided with a full listing of all inflected forms and other morphological variation. Every word form is marked for part of speech (using Penn TreeBank notation). Most single-word forms include a representation of IPA pronunciation. BESL covers both British and American English, and other spelling variants, with cross-references between corresponding forms. Each lemma is graded on a scale between 1 and 9 to indicate frequency, based on corpus evidence. Lemmas are also classified by domain, where appropriate (e.g. Computing, Religion). Obscene or offensive lemmas are marked using a 2-grade system. Proper name lemmas in BESL include personal names, surnames, place names, and brand names. BESL is provided in XML.

	ELRA members	Non-members
For research use	7,000 Euro	10,000 Euro

ELRA-L0059 Offensive Word Filter 1

Oxford University Press has developed two lists of offensive words and expressions, specifically developed for filter applications in the contexts of web pages and email. Each list features a grading system describing vocabulary type and offensive strength for each term, plus collocational information to help identify the terms in context.

Coverage: 4500 words and expressions; 10-category classification system; UK and US usage covered, plus other world English

Features: graded by class (offensive/vulgar), and type (racist, sexist etc); rated by strength (high/moderate/mild); part of speech included; morphological status marked (standalone/fixed collocation etc); collocational information included; practical screening recommendation

Format: tab-delimited ASCII

File Size: 262kB

	ELRA members	Non-members
For research use	4,000 Euro	5,000 Euro

ELRA-L0060 Offensive Word Filter 2

Oxford University Press has developed two lists of offensive words and expressions, specifically developed for filter applications in the contexts of web pages and email. Each list features a grading system describing vocabulary type and offensive strength for each term, plus collocational information to help identify the terms in context.

Coverage: over 2000 words and expressions; 13-category classification system; US and UK usage covered

Features: graded by category/subcategory (eg abusive/sexist etc); rated by strength (extreme/moderate/mild); collocational information included; regional usage/source labelling; glosses for obscure senses

Format: Excel spreadsheet

File Size: 237 kB.

	ELRA members	Non-members
For research use	2,000 Euro	2,500 Euro

ELRA-L0061 The Oxford Spanish Dictionary

The highly-acclaimed Oxford Spanish Dictionary, described by John Butt in the TLS (Times Literary Supplement) as "indispensable for all serious Hispanists", provides an authoritative, up-to-date guide to world Spanish. It is the only Spanish dictionary to present the full wealth of Spanish from both sides of the Atlantic, with coverage of 24 varieties of Spanish as it is written and spoken throughout the Spanish-speaking world. There are thousands of real, authentic example sentences carefully selected to illustrate the full range of meanings and typical contexts.

Coverage: 300,000 words and phrases; 500,000 translations; 24 regional varieties of Spanish included

Features: thousands of example sentences, from written sources and transcripts of real speech; up-to-the-minute coverage of general, scientific, literary, and technical vocabulary; comprehensive coverage of idioms

Format: Available in XML or SGML

	ELRA members	Non-members
For research use	6,125 Euro	8,750 Euro

ELRA-L0062 French Source Lexicon

Oxford University Press currently holds extensive databases of morphological and phonetic data for Spanish, French, and Italian. Each headword lemma is provided with a full listing of its possible syntactic forms and spelling variants, along with information on their relationship to the headword form. In addition, a representation of the IPA pronunciation is given for every form. There is also information on domains in which the headwords are used, e.g. Computing, Engineering, Zoology. The lexicon is provided in SGML.

Coverage: over 90,000 headwords/lemmas, over 400,000 wordforms, over 1,000 abbreviations, over 35,000 proper nouns

Features: clear indication of preferred orthographic forms, with cross-references from variants, exceptional coverage of place names, both French and worldwide, high-profile French sources

Format: Available in SGML

	ELRA members	Non-members
For research use	7,000 Euro	10,000 Euro

ELRA-T0368 Multilingual Wordbank

Alongside its world-renowned range of Bilingual Dictionaries, Oxford University Press holds both word and phrase translation glossaries designed for the travel/handy-reference market. These feature translations of core vocabulary from English into French, German, Italian, Spanish, and Portuguese, plus full coverage of local variations in American English, Latin American Spanish, and Brazilian Portuguese.
Every word and phrase is given a frequency ranking, which can be used as a guide to user levels. In addition, all translations in the Wordbank are provided along with appropriate part of speech and gender information, while the Phrasebank also has gender information where relevant to the syntax.
Coverage: 17,500 core terms (per language), 9 languages covered (including regional variants): UK English, US English, French, German, Italian, European Spanish, Latin American Spanish, European Portuguese, Brazilian Portuguese
Features: frequency rating for user level, gender + part of speech information
Format: tab-delimited text

	ELRA members	Non-members
For research use	4,000 Euro	5,000 Euro

ELRA-T0369 Multilingual Phrasebank

The Phrasebank has phrases organized under 9 different topics, many of which are further subdivided. It is presented in a compressed format, with substitutable elements bracketed, and one or several alternatives included within the entry, reducing storage space wasted due to repetition of common material. The compression is extended further by reference to "template" sets of common terms, e.g. Days of the Week, Parts of the Body, allowing a base phrase to be combined with up to 100 different terms.
Coverage: 3,000 base phrases (per language), 5,000 expanded phrases (excluding templates); 20,000 expanded phrases (including templates), 9 languages covered (incl regional variants): UK English, US English, French, German, Italian, European Spanish, Latin American Spanish, European Portuguese, Brazilian Portuguese
Topics: Diversions, Eating Out, Entertainment, Hotel, Money, Problems, Tourist, Transport, Miscellaneous
Templates: Body Parts, Cardinal Numbers, Clock Time, Colours, Currencies, Dates, Day of Week, Languages, Months, Nationalities, Ordinal Numbers, Places
Features: frequency rating for user level, gender information as required, compressed format reducing storage space for phrase variants, 9 topics/15 subtopics, extension of selected phrases using 12 template lists
Format: tab-delimited text for phrases; Excel spreadsheets for template lists.

	ELRA members	Non-members
For research use	4,000 Euro	5,000 Euro

ELRA-T0370 Dictionary of Law

This fully up-to-date edition takes account of recent changes in UK legislation. Over 4,000 entries define and explain the major terms, concepts, processes, and the organization of the English legal system. It features authoritative and up-to-date articles which have been written by practising and academic lawyers. New entries cover the Woolf reforms, human rights law, as well as family law, central and local government, and international law. The dictionary is provided in XML.

	ELRA members	Non-members
For research use	4,000 Euro	5,000 Euro

ELRA-T0371 Dictionary of Medicine

Over 10,000 clear and concise entries cover all major medical and surgical specialities. The dictionary reflects recent developments in the medical field, covering new drugs in clinical use, as well as new advances in genetics, infertility treatment, cancer, organ transplantation, and BSE. The dictionary is provided in XML.

	ELRA members	Non-members
For research use	4,000 Euro	5,000 Euro

ELRA-E0002: TC-STAR Evaluation Package - ASR English

The TC-STAR Evaluation Package - ASR English includes the material used for the TC-STAR 2005 Automatic Speech Recognition (ASR) first evaluation campaign for the English language. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of this evaluation package is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

ELRA-E0003: TC-STAR Evaluation Package - ASR Spanish

The TC-STAR Evaluation Package - ASR Spanish includes the material used for the TC-STAR 2005 Automatic Speech Recognition (ASR) first evaluation campaign for the Spanish language. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of this evaluation package is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

ELRA-E0004: TC-STAR Evaluation Package - ASR Mandarin Chinese

ASR Mandarin Chinese includes the material used for the TC-STAR 2005 Automatic Speech Recognition (ASR) first evaluation campaign for the Mandarin Chinese language. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of this evaluation package is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

ELRA-E0005: TC-STAR Evaluation Package - SLT English-to-Spanish

ASThe TC-STAR Evaluation Package - SLT English-to-Spanish includes the material used for the TC-STAR 2005 Spoken Language Translation (SLT) first evaluation campaign for English-to-Spanish translation. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of this evaluation package is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.-

ELRA-E0006: TC-STAR Evaluation Package - SLT Spanish-to-English

The TC-STAR Evaluation Package - SLT Spanish-to-English includes the material used for the TC-STAR 2005 Spoken Language Translation (SLT) first evaluation campaign for Spanish-to-English translation. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of this evaluation package is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself

ELRA-E0007: TC-STAR Evaluation Package - SLT Chinese-to-English

The TC-STAR Evaluation Package - SLT Chinese-to-English includes the material used for the TC-STAR 2005 Spoken Language Translation (SLT) first evaluation campaign for Chinese-to-English translation. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of this evaluation package is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

PRICES FOR TC-STAR EVALUATION PACKAGES

Prices per package

	ELRA members	Non-members
For research use	500 Euro	750 Euro

Special prices for a combined purchase TC-STAR Evaluation Packages

- **ASR Suite (E0002 + E0003 + E0004):**

	ELRA members	Non-members
For research use	1,200 Euro	1,800 Euro

- **SLT Suite (E0005 + E0006 + E0007):**

	ELRA members	Non-members
For research use	1,200 Euro	1,800 Euro

- **ASR + SLT Suites (E0002 + E0003 + E0004 + E0005 + E0006 + E0007):**

	ELRA members	Non-members
For research use	2,000 Euro	3,000 Euro