The ELRA Newsletter

July - December 2007

Vol.12 n.3&4

Contents

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

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,

This is the last issue of 2007 and we would like to quickly highlight a number of topics on which ELRA focussed during this year.

The major event for ELRA is the organisation of **LREC 2008**, the sixth edition in the series of the Language Resources and Evaluation Conference launched by ELRA with the support of a large number of active players in the field. Therefore, ELRA and ELDA have been strongly involved in the preparation of the 6th edition that will take place in Marrakech, Morocco, from May 26th to June 1st, 2008.

Strong focus has also been put on the **Universal Catalogue** for which we also devoted time and manpower trying to enrich it with as many resources as possible while also maintaining existing ones with updated descriptions and contacts.

Over the past few months, ELRA and ELDA have continued to work on their regular activities, such as the tasks carried out within the various committees (VCom, PCom) and within a number of projects.

As for this newsletter, it contains:

- **The UK: an overview of some current work in HLT**, which gives an overview of the activities taking place in the UK with regard to HLT.
- A contribution describing **TwNC, the Twente News Corpus, a multifaceted corpus for Dutch**.

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

- *Monolingual Lexicons from the general domain:*
  - Polderland Dutch Lexicon of Abbreviations and Acronyms (ELRA-L0076)
  - Polderland Dutch General Lexicon (ELRA-L0077)
  - Polderland Dutch Lexicon of Names (ELRA-L0078)
  - Polderland Dutch Lexicon of Business Terminology (ELRA-L0079)
  - Polderland Dutch Lexicon of Legal Terminology (ELRA-L0080)
  - Polderland Dutch Lexicon of Medical Terminology (ELRA-L0081)
  - Polderland Dutch Lexicon of Social Terminology (ELRA-L0082)
  - Polderland Dutch Lexicon of Technical Terminology (ELRA-L0083)
  - Macedonian Morphological Lexicon (MACPLEX) (ELRA-L0084)
- *Phonetic lexicons from the LC-STAR European-funded project:*
  - LC-STAR German Phonetic Lexicon (ELRA-S0245)
  - LC-STAR German Phonetic Lexicon in the Touristic Domain (ELRA-S0246)
  - LC-STAR Standard Arabic Phonetic Lexicon (ELRA-S0247)
  - LC-STAR English-German Bilingual Aligned Phrasal Lexicon (ELRA-S0248)
  - LC-STAR Finnish Phonetic Lexicon (ELRA-S0255)
  - LC-STAR Mandarin Chinese Phonetic Lexicon (ELRA-S0256)
  - LC-STAR English-Finnish Bilingual Aligned Phrasal Lexicon (ELRA-S0257)
- *Speech Microphone resources from the Speecon European-funded project:*
  - Japanese Speecon Database (ELRA-S0244)
  - Dutch from Belgium Speecon Database (ELRA-S0265)
  - Dutch from the Netherlands Speecon Database (ELRA-S0266)
  - Danish Speecon Database (ELRA-S0267)
- *Speech Microphone resources from the TC-STAR European-funded project:*
  - TC-STAR English Training Corpora for ASR: Transcriptions of EPPS Speech (ELRA-S0249)
  - TC-STAR English-Spanish Training Corpora for Machine Translation: Aligned Final Text Editions of EPPS (ELRA-S0250)
  - TC-STAR English Training Corpora for ASR: Recordings of EPPS Speech (ELRA-S0251)
  - TC-STAR Spanish Training Corpora for ASR: Recordings of EPPS Speech (ELRA-S0252)
  - TC-STAR English Test Corpora for ASR (ELRA-S0253)
  - TC-STAR Spanish Test Corpora for ASR (ELRA-S0254)
- *Speech Telephone Database from the Orientel European-funded project:*
  - Orientel United Arab Emirates MCA (Modern Colloquial Arabic) (ELRA-S0258)
  - Orientel United Arab Emirates MSA (Modern Standard Arabic) (ELRA-S0259)
  - Orientel English as spoken in the United Arab Emirates (ELRA-S0260)
- *Speech Telephone from the SpeechDat(E) European-funded project:*
  - Hungarian SpeechDat(E) Database (ELRA-S0261)
- *Speech Telephone based on the SpeechDat project conventions:*
  - SpeechDat Catalan FDB Database (ELRA-S0243)
- *Speech Telephone Database from the SALA European-funded project:*
  - SALA II US English Database (ELRA-S0242)
  - SALA II Portuguese from Brazil Database (ELRA-S0262)
  - SALA II Spanish from Colombia Database (ELRA-S0263)
  - SALA II US Spanish West Database (ELRA-S0264)
- *Evaluation Packages:*
  - AURORA-5 (AURORA-CD0005)
  - TC-STAR 2006 Evaluation Package - End-to-End (ELRA-E0031)
  - TC-STAR 2007 Evaluation Packages:
    - ASR English (ELRA-E0025)
    - ASR Spanish - CORTES (ELRA-E0026-01)
    - ASR Spanish - EPPS (ELRA-E0026-02)
    - ASR Mandarin Chinese (ELRA-E0027)
    - SLT English-to-Spanish (ELRA-E0028)
    - SLT Spanish-to-English - CORTES (ELRA-E0029-01)
    - SLT Spanish-to-English - EPPS (ELRA-E0029-02)
    - SLT Chinese-to-English (ELRA-E0030)
    - End-to-End (ELRA-E0032)

Once again if you would like to join ELRA and benefit from its services (that are summarized at www.elra.info), please contact us.

Bente Maegaard, President

Khalid Choukri, CEO

The UK: an Overview of Some Current Work in HLT

Martin Wynne

E-Science and the Human Language Technologies

The e-Science vision sees beyond a fragmented field of diverse resources and tools, and isolated researchers, to a world of shared data, interoperable tools and collaborative research, underpinned by the next generation of networked computing power. Many researchers and funders in the UK now see the potential for the language resources and technologies as part of the e-Science agenda.

In a significant recent development, the UK government's Office for Science and Innovation (OSI) published a report 'Developing the UK's e-Infrastructure for Science and Innovation'. This report recognises that a national e-infrastructure for research provides a vital foundation for the UK's science base, supporting not only rapidly advancing technological developments, but also the increasing possibilities for knowledge transfer and the creation of wealth. While we are currently a long way away from constructing such an infrastructure for the language technologies, UK researchers are playing leading roles in international developments.

The CLARIN project is aiming to build a pan-European infrastructure to support the creation and use of language resources and tools in research across the humanities and social sciences. A wide network has developed of researchers in the UK involved with building the CLARIN infrastructure, and interested more widely in e-infrastructure issues, co-ordinated by the Oxford Text Archive.

One of the grand challenges of e-Science is to develop methods to deal with the data deluge. Tens of thousands of scientific papers are published each week, and without sophisticated tools, researchers will be unable to monitor and understand developments in their field. Text mining (TM) techniques can find knowledge hidden in text and to present it in a concise form, and as text mining matures, it will increasingly enable researchers to collect, maintain, interpret, curate, and discover knowledge needed for research and education. The National Centre for Text Mining (NaCTeM) is playing a critical role in ensuring that UK researchers are aware of and have access to effective TM solutions, and are able to exploit their capabilities to the full. NaCTeM is a major initiative which has emerged in recent years and is the first publicly-funded text mining centre in the world. NaCTeM is operated by the University of Manchester with close collaboration with the University of Tokyo.

E-Science cuts across subject boundaries and organisational structures. Many projects, like those at NaCTeM, deploy language tools to aid research in the physical sciences. Language technologies are also being deployed in the social sciences, and in the arts and humanities.

The Economic and Social Research Council e-Social Science node 'Understanding New Forms of Digital Records for e-Social Science', in collaboration with the School of English Studies at the University of Nottingham has recently completed a project to explore language in relation to gestures captured in multi-modal corpora.

The Arts and Humanities e-Science Support Centre has conducted an e-Science Scoping Study, and one of the subject areas covered was Linguistics. While drawing attention to exciting and innovative work in this area, the survey also identified barriers to effective research where linguistics researchers need increased support in the areas of data acquisition and management, data annotation, data access and data retrieval. The importance of developing a research infrastructure was clearly identified.

Alongside these initiatives in the social sciences and humanities, much research in speech and language technologies in the UK continues as part of the remit of the Engineering and Physical Sciences Research Council, with current projects including work on speech production, recognition and synthesis, anaphora resolution, computational semantics, metaphor, gesture, second language fluency, computer-mediated communication, and more.

The Emerging West Midlands

While much impressive work in speech and language technologies continues at the well-known centres of expertise, some of the most interesting new research is also taking place in less well-known institutions. Two examples are examined here, at Birmingham City University (BCU) and University of Wolverhampton.

Antoinette Renouf is leading several new projects at the Research and Development Unit for English Studies at Birmingham City University (formerly University of Central England in Birmingham). An investigation into "Repulsion" is extending work on collocation by analysing corpus evidence to look for evidence of a tendency for words not to co-occur. This could have profound impact on our understanding of the creation of meaning and style in language, and serve as a tool to identify errors and evaluate suggested choices in drafts by writers of text and international users of English.

Work at BCU also continues on the WebCorp Linguist's Search Engine, a large-scale specialist search engine which is being integrated with a range of language-analysis and output-formatting tools. Linguists, language engineers, teachers and students are increasingly trying to use the World Wide Web as a source of linguistic information to supplement the information on language use in existing dictionaries and text collections, and this project aims to help them to do so in more sophisticated and accurate ways.

A few miles from BCU in the English West Midlands, the Research Group in Computational Linguistics, University of Wolverhampton has a longer history, founded in 1998 by Ruslan Mitkov and producing cutting edge research in a variety of areas of computational linguistics. The Wolverhampton group is well known for its contribution to the field of anaphora and coreference resolution, and in the last few years the research group has also successfully completed research in automatic summarisation, and automatic translation. A current project, QALL-ME, with European funding, is developing a question answering system for multilingual and multimodal environments. The project has proposed a new question answering method based on

textual entailment and produced a first prototype that works in the domain of cinema and movies. In addition, the emphasis of the project is on producing an open architecture for question answering based on web services which allows easy integration of processing modules.

Alongside the long-established work in corpus and computational linguistics at the University of Birmingham, and recent relaunch of corpus research at Aston University, the West Midlands are an exciting place for HLT in 2008.

New Developments in XML

The recent, long-awaited release of an XML version of the British National Corpus has led to an increased uptake in use of this resource. As well as serving as an invaluable large dataset of recent British English usage, it is becoming useful as a historical corpus, as linguists and social historians start to examine changes in the spoken and written language in the period since the corpus was compiled in the late twentieth century. It is anticipated that this release of the corpus in XML will make it more easily usable with the latest applications, as well as ensuring its ongoing sustainability and usefulness as a historical, synchronic corpus.

The latest version of the guidelines of the Text Encoding Initiative (TEI), known as P5, was released in November 2007. While the TEI is a truly international community, and the guidelines are developed with input from around the globe, the European Editor of the guidelines is Lou Burnard of Oxford University Computing Services, and key sections of the guidelines and important associated XML technologies have been developed by UK researchers. The TEI Guidelines for Electronic Text Encoding and Interchange define and document a markup language for representing the structural and conceptual features of texts. They focus on the encoding of documents in the humanities and social sciences, and in particular on the representation of primary source materials for research and analysis, but it is expected that they will also have a big impact on work in NLP around the world, and an important reference point for standards in developing language resources.

Links

1) National e-Science Centre: http://www.nesc.ac.uk
2) e-Science and Linguistics Scoping Survey: http://ahds.ac.uk/e-science/documents/Rayson-report.pdf
3) CLARIN: http://www.clarin.eu
4) National Centre for Text Mining: http://www.nactem.ac.uk
5) Economic and Social Research Council: http://www.esrc.ac.uk/ESRCInfoCentre/index.aspx
6) Arts and Humanities e-Science Support Centre: http://www.ahessc.ac.uk/ahessc-home
7) Engineering and Physical Sciences Research Council: http://www.epsrc.ac.uk/default.htm
8) EPSRC 'Grants on the Web': http://gow.epsrc.ac.uk
9) Multimedia Information Systems: http://mmis.doc.ic.ac.uk
10) Research and Development Unit for English Studies at Birmingham City University: http://rdues.bcu.ac.uk
11) Research Group in Computational Linguistics, University of Wolverhampton: http://clg.wlv.ac.uk
12) British National Corpus: http://www.natcorp.ox.ac.uk
13) Text Encoding Initiative: http://www.tei-c.org

Martin Wynne
Head of the Oxford Text Archive
University of Oxford
martin.wynne@oucs.ox.ac.uk

TwNC: a Multifaceted Dutch News Corpus

Roeland Ordelman, Franciska de Jong, Arjan van Hessen, Hendri Hondorp

Introduction

This contribution describes the Twente News Corpus (TwNC), a multifaceted corpus for Dutch that is being deployed in a number of NLP research projects among which tracks within the Dutch national research programme MultimediaN, the NWO programme CATCH, and the Dutch-Flemish programme STEVIN(1). The development of the corpus started in 1998 within a predecessor project DRUID and has currently a size of 530M words. The text part has been built from texts of four different sources: Dutch national newspapers, television subtitles, teleprompter (autocues) files, and both manually and automatically generated broadcast news transcripts along with the broadcast news audio. TwNC plays a crucial role in the development and evaluation of a wide range of tools and applications for the domain of multimedia indexing, such as large vocabulary speech recognition, cross-media indexing, cross-language information retrieval, etc. Part of the corpus was fed into the Dutch written text corpus in the context of the Dutch-Belgian STEVIN project D-COI that was completed in 2007. The sections below will describe the rationale that was the starting point for the corpus development; it will outline the cross-media linking approach adopted within MultimediaN, and finally provide some facts and figures about the corpus.

(1) Relevant links are listed at the end of this article.

Text Corpora and the Development of Semantic Access to Multimedia via Natural Language

To tackle the challenge handed in by the ever increasing volumes of multimedia content being created and stored on the one hand, and the promising steps made in multimedia analysis on the other hand, it is tempting to put the focus on the advancement of image analysis and its integration with recently gained insights from relevant domains such as knowledge extraction and semantic web technology. Still, the rationale within programmes such as

MultimediaN is that it would be grossly erroneous to ignore the available insights in the successful role that can be played by the modality for which very matured analysis frameworks exist: natural language.

As is widely acknowledged, the exploitation of linguistic content in multimedia archives can boost the accessibility of multimedia archives enormously. Already in 1995, Brown et al(2) demonstrated the use of subtitling information for retrieval of broadcast news videos, and in the context of TRECVID a common feature of the best performing video retrieval systems is that they exploit speech transcripts. Of course the added value of linguistic data is limited to video data containing textual and/or spoken content, or to video content with links to related textual documents, e.g., subtitles, manually generated transcripts, etc. But, where available, linguistic content can play a crucial role bridging the semantic gap between media features and user needs.

Depending on the resources available within an organization that administers a media collection, the amount of detail of the metadata and their characteristics may vary. Large national audiovisual institutions annotate at least descriptive metadata: titles, dates and short content descriptions. However, many multimedia archiving institutes often do not have the resources to apply even some basic form of archiving. In order to still allow the conceptual querying of video content, collateral textual resources that are closely related to the collection items can be exploited. They can be either available because they play a role in the production or broadcast process, or they can be generated via speech recognition. Collateral data can be exploited to (i) produce highly accurate, automatic indexes in an affordable way, (ii) tune speech and language processing tools to the focus domain, and (iii) enhance presentation of video retrieval search results by adding extra layers of information.

A well known example of such a collateral textual resource is subtitling information for the hearing-impaired (e.g., CEEFAX pages 888 in the UK) that is available for the majority of contemporary broadcast items, and in any case, for news programs. Subtitles contain a nearly complete transcription of the words spoken in video items and can be easily linked to the video by using the time-stamps that come with the subtitles. Textual sources that can play a similar role are teleprompter files: the texts read from screen by an anchor person (also referred to as auto-cues). Also outside the broadcast sector, collateral materials that somehow match the speech in a collection can be found. A collection of recorded lectures may have presenter notes attached to it, speeches may be accompanied with the written text version, and for meeting recordings there may be minutes available, or at least an agenda.

In absence of (or lack of access to) such error-free texts, there is always the possibility to use automatic speech recognition (ASR) for the generation of transcripts. Relatively limitedly referenced, however, is the exploitation potential for textual content to complement speech transcripts. In all the examples mentioned above the time stamps in the sources are crucial for the creation of a textual index into video. In collateral text sources, the available time-labels are not always fully reliable and outside the news broadcast domain they will often be absent. The resynchronization or labelling of the text with time-codes is called the 'alignment' of text and speech, a well-known procedure used frequently in ASR, for example, when training acoustic models. The alignment of collateral data holds for surprisingly low text-speech correlation levels, especially when some additional trickery is applied.

Recent years have shown that large vocabulary speech recognition can successfully be deployed for creating multimedia annotations allowing the conceptual querying of video content. When the collateral data only correlates with the speech on the topic level, full-blown speech recognition must be called in, using the collateral data as a strong prior ('informed speech recognition') or source for extensive domain tuning via the language model. Alternatively, the collateral data can be used for relevance feedback during search. Also the out-of-vocabulary rate could be decreased: if a (non-perfect) ASR transcript is used as the basis for a search of related text, and the terms referring to named entities in the most similar texts are fed into the language models, a second run of the ASR could yield improved recognition results.

Ideally one would not only synchronize audiovisual material with content that approximates the speech in the data, but take even one step further and exploit any accessible text including open source titles and proprietary data (e.g., trusted web pages and newspaper articles). In the context of meetings for example, usually an agenda, documents on agenda topics and CVs of meeting participants can be obtained and linked to the media repository.

Finally, there is of course also the possibility to use an audio fragment as a query for textual documents. Via a transcription of an audio query, related text can be identified. An obvious application domain for this option is, again, news. But it works also in other domains than news, e.g., oral history archives, meeting or lecture recordings, audio blogs, digital storytelling, etc.

The Structure of the Twente News Corpus

The original goal for starting the development of the Twente News Corpus (TwNC) was to collect data for the training of language models and acoustic models to be incorporated into a system for large vocabulary speech recognition for Dutch to be deployed in the broadcast news domain and, also, as a baseline system in other domains that lack large amounts of example data (e.g., cultural heritage data as we encounter in the Dutch CHoral project). The focus on news was given in by the size of the datasets available for this domain, and by the focus on news as target at many other research groups. News is a target domain for corpus development, for search applications and for speech technology.

Several requirements come from this type of deployment for a text corpus. They pertain to formatting, encoding, size and balancing, for example. TwNC text data has been formatted as XML and the encoding chosen is utf-8. Balancing is reached by selecting four different source types: newspaper text, autocue files (telepromp-

(2) M. G. Brown, J.T. Foote, G.J.F. Jones, K. Sparck Jones, and S. J. Young. Automatic Content-based Retrieval of Broadcast News. In Proceedings of the third ACM international conference on Multimedia, pages 35-43, San Francisco, November 1995. ACM Press.

ter text), subtitling files and manually generated transcripts. The current corpus size is approximately 500 million words of text and about 800 files of broadcast news audio. In the remainder of this section we will describe the types in more detail.

Newspaper data

One of the largest publishers in the Dutch language region, PCM publishers, have donated content on a daily basis (via ftp) from six national newspapers. UT was given access to two years ('94-'95) of content from *NRC Handelsblad* and *Trouw*. Since 1999 also content from four other newspaper titles is made available: *Algemeen Dagblad*, *Volkskrant*, *Parool* and *NrcNext*. There are even a few years for which content is available from magazines such as *Vrij Nederland* and *HP De Tijd*, and soon also *Groene Amsterdammer*. In Table 1 and Table 2 the statistics of the newspaper data are listed. Daily newspaper feed is not just helping to enlarge the corpus, it also facilitates the updating of the broadcast news vocabulary, or the daily production of word occurrence statistics and predictions such as illustrated in Figure 1. A number of research groups have explored parts of the newspaper corpus for tasks such as measuring lexical variation, paraphrase identification learning, and extracting hypernym-hyponym relations.

Subtitles

Since 1998 we have been capturing subtitling for the hearing impaired of broadcast news shows that are normally projected on the television screen in the Netherlands via teletext page 888 (CEEFAX pages 888 in the UK) but can also be accessed directly using a TV-card.

Due to a minimum of available space for subtitles on a screen, the number of words in the subtitles is cut down drastically compared to what was actually said. Although phrases are often mixed up completely in an attempt to say the same with less and often other words, subtitles provide an excellent information source for automatic indexing. Moreover, within the Dutch broadcast news autocue files, subtitle topics are marked which is very interesting from a low cost indexing perspective. Finally, the subtitles have time information attached to them referring to the exact time the subtitle was projected on the screen. The delay with respect to the speech differs depending on whether the subtitles are

Year	Mwords
1994	34
1995	33
1999	63
2000	82
2001	70
2002	70
2003	34
2004	40
2005	36
2006	41
2007	37
2008	7
Total	**547**

Table 1: Number of newspaper words (in millions) per year

Newspaper/Magazine	NPs	Mwords
NRC Handelsblad	2913	156
Volkskrant	2392	137
Algemeen Dagblad	2375	102
Trouw	2020	85
Parool	1475	58
NrcNext	169	5
Dordtsch Dagblad	117	2
HP De Tijd	21	1
Vrij Nederland	18	1
Total	**11500**	**547**

Table 2: Number of newspapers and words per newspaper/magazine since 1994

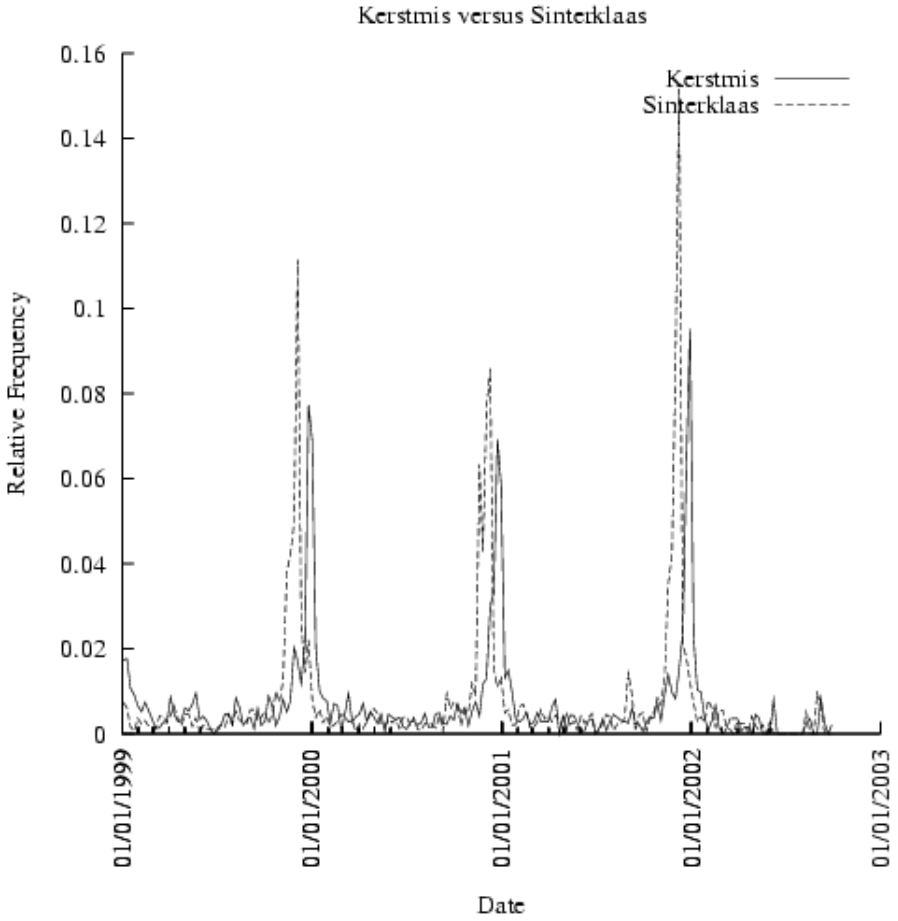


Figure 1: Word occurrence statistics of the words Christmas and Santa-Claus from 1999-2003

generated live or not. The broadcast news retrieval demonstrator that is running at the University of Twente shows how both speech recognition and subtitle information can be deployed for indexing a news show.

The captured teletext subtitles have been converted to a suitable XML format that includes the topic boundaries and time information. In total, the Twente News Corpus has 2.3 M words of broadcast news subtitles. For a subset of this collection also the audio of the broadcast is available.

Teleprompter files

The third type of news related text data we collected consists of *teleprompter* files also referred to as *autocues*: the texts the newsreaders read from the teleprompter. The autocues have been kindly provided between 1998 and 2005 by the Dutch National Broadcast Foundation (NOS), the producer of the 8 o'clock news show (8 Uur Journaal). The autocues are almost an exact representation of the speech from the newsreaders but lack of course spontaneous utterances and live commentaries ('pseudo-live' commentaries are often included). There is topic information and (relative) time information available in the files. The RTF format autocues were converted to XML and have a total word count of 3.3 M words.

Transcripts and audio

For the purpose of training acoustic models for a broadcast news speech recognition system, 26 broadcast news shows were recorded and manually annotated on the word level. From 2005 onward, both audio and subtitles from the 8 o'clock news shows that were recorded and indexed in our broadcast news retrieval demonstrator were preserved. These data pairs (currently some 800) can then for example be used for the partly unsupervised training of acoustic models. With a non optimised routine we automatically extracted about 60 hours of training data consisting of a lot of small audio segments with aligned transcripts (sequence of 3 words minimum) from 279 news shows (2006). UT is currently also processing other years and intends to make the acoustic models that are trained with the data available (see also 'Tools' below).

	with transcripts	with subtitles	with autocues
Audio (828)	26	801	30

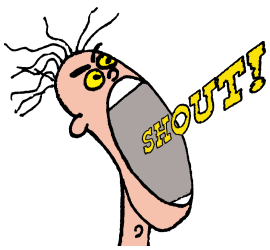
Availability Conditions

The newspaper content is made available for use by researchers under the condition that they do not publish any summaries, analyses or interpretations of the linguistic characteristics that can lead to extraction of reconstruction of the original content. UT is allowed to redistribute portions of the data under strict licence agreements. Currently, access to the 1999-2002 data can be licensed to individual research groups. The 1994-1995 data was redistributed among the participants of the evaluation campaign within CLEF (Cross-Language Evaluation Forum). Recently, the 1999-2004 data was made available to participants of the Dutch STEVIN speech recognition benchmark evaluation N-BEST, specifically for purposes of language model research.

Source	Amount
Newspapers	547Mw
BN Subtitles	2.3Mw
BN Autocues	3.3Mw
BN Audio	828
BN Transcripts	26
BN Alignments	60+ hours

Table 3: Overview of the TwNC

Tools



Along with the corpus itself UT can provide tools developed for text normalisation purposes and speech recognition: the UT Text Normalisation Toolkit (UT-TNT), that was partly developed in the MultimediaN project and the STEVIN project SPRAAK, and the SHoUT speech recognition toolkit developed in MultimediaN. SHoUT Acoustic models that are trained using the automatic alignment procedure will be made available at a later stage.

Links

1) Twente News Corpus:
http://hmi.ewi.utwente.nl/twnc

2) MultimediaN:
http://www.multimedian.nl

3) NWO programme CATCH:
http://www.nwo.nl/catch (in Dutch only)

4) STEVIN:
http://www.taaluniversum.org/stevin (in Dutch only)

5) DRUID Project : http://hmi.ewi.utwente.nl/Projects/druid.html

6) STEVIN projects D-COI & SPRAAK:
http://hmi.ewi.utwente.nl/project/STEVIN

7) Choral Project:
http://hmi.ewi.utwente.nl/choral

8) Broadcast news retrieval demonstrator:
http://hmi.ewi.utwente.nl/showcases/broadcast-news-demo

9) Cross-Language Evaluation Forum (CLEF):
http://www.clef-campaign.org

10) Shout Speech Recognition Toolkit:
http://www.vf.utwente.nl/~huijbreg/shout/index.html

11) TRECVID:
http://www-nlpir.nist.gov/projects/trecvid

Roeland Ordelman, Franciska de Jong, Arjan van Hessen, Hendri Hondorp

University of Twente (UT)
Department of Electrical Engineering, Mathematics and Computer Science
Human Media Interaction Group
P.O. Box 217, 7500 AE Enschede
The Netherlands

http://hmi.ewi.utwente.nl

ordelman@ewi.utwente.nl
fdejong@ewi.utwente.nl
hessen@ewi.utwente.nl
hendri@cs.utwente.nl

ELRA's Activities in HLT Evaluation

ELRA's Mission

ELRA's initial mission was to set up a centralized not-for-profit organization for the collection, distribution, and validation of speech, text, terminology resources and tools. In order to play this role of a central repository, ELRA had to address issues of a different nature such as technical and logistic problems, commercial issues (prices, fees, royalties), legal issues (licensing, Intellectual Property Rights, …), information dissemination (to act as a clearing house). This mission is tuned from time to time to anticipate future requirements. As of today, this can be reflected by the following tasks:

- The identification of useful Language Resources.
- The handling of legal issues related to the availability of Language Resources.
- The Language Resource distribution activities and pricing policy.
- The validation and quality assessment of Language Resources.
- The commission of the production of needed Language Resources & market watch.
- The information dissemination, promotion and awareness.
- Last but not least, the supply of the evaluation services to the HLT community.

HLT Evaluation Activities

For any HLT research effort to be successful, it is essential that it be assessed through rigorous evaluations of the developed technologies. This allows performance benchmarking and a better understanding of possible limitations and challenging conditions. Since 2000, ELDA, as ELRA's operational body, had an objective to design and validate evaluation packages for several Human Language Technologies. An ELDA evaluation package (or Evaluation Kit) comprises the following items:

1) An evaluation protocol specification, including specification of the task to be performed by the systems being evaluated, metrics, data representation formalisms, and the relevant documentation.

2) Development data representative of the task and in sufficient amount to enable a full validation of the evaluation protocol.

3) The test data that will be used to score system performance.

4) All the software tools required to run an evaluation campaign implementing the protocol defined in 1), i.e. format standardization and validation tools, measuring tools, result presentation tools, data server, storing and communication tools, etc.

These evaluation packages had to be made available for organizing large evaluation campaigns, involving all key players from laboratories developing technologies targeted for evaluation. The evaluation packages also had to be made commercially available upon request for government agencies or industries wishing to organize other evaluation campaigns. Finally, these packages are distributed for industrial or public research entities wishing to evaluate a technology (possibly the one they develop) and compare it to the state-of-the-art.

In many cases, ELDA also helps to provide data for system training and, since 2006, it provides an on-line evaluation platform for several technologies (web-service and/or UIMA-based platform) to avoid the installation of scoring tools at each site.

In order to achieve such goal, ELDA has established an Evaluation Department that takes care of assessing and benchmarking Human Language Technologies both within R&D projects and for customers. In order to do so, ELDA has joined efforts with some of the largest consortia involved in HLT development and it has managed to ensure that the consortia capitalize on the evaluation work through the packaging of all needed pieces to carry out similar initiatives afterwards. A new paradigm refers to this task as the "project exit strategy". Such strategy ensures that the availability of the "evaluation package" as described above (the full documentation, definition and description of the evaluation methodologies, protocols and metrics, alongside the data sets and software scoring tools) is an essential part of each project. An evaluation package can be conditioned so that it can be distributed through ELRA's Catalogue. This allows any organization to reproduce one of the technology evaluations that were conducted during the project Evaluation Campaign. The exploitation of this outcome is one of the achievements of the project.

ELRA's Focus on Evaluation

The ELSE project (Evaluation in Language and Speech Engineering, 1999) conducted a study on the possible implementation of HLT evaluations in Europe and compared them to other initiatives conducted in the USA and Japan. Among the findings of this project we can quote the need for comparative evaluations conducted by a truly European infrastructure that would ensure long-term availability of expertise and resources so as to avoid the loss that occurs when projects are funded through R&D programs for a short period of time. The ELSE project also highlighted how important this was for the developers that benefit indirectly from evaluation through the acquisition of the complete evaluation toolkits and by-product data that become available afterwards but also through the knowledge sharing that takes place systematically in the post-campaign workshops during which experts compare approaches and techniques used by each system.

Within these evaluation activities, ELDA has participated in a large number of projects and initiatives which have helped reinforce its expertise in the area and have supported the development of HLT Evaluation in Europe. Among these projects we will mention just a few here to highlight the huge European investment and the crucial need to ensure a serious return on investment through the exploitation of such packages but also through the support of the ELDA infrastructure to become self-sustainable.

A hot topic these days is Machine Translation technology, including Speech to Speech translation systems. Through its involvement in the FP6 project TC-STAR, ELDA has contributed to the evaluation of speech recognition systems, machine translation, and speech synthesis systems. In addition, ELDA conducted end-to-end evaluations and compared the achievements of TC-STAR systems with the work

of human interpreters for English and Spanish.

TC-STAR

One of ELDA's tasks within this project was the collection and annotation of very large sets of spoken multilingual corpora and the corresponding written corpora that was used to train the systems. ELDA has also been in charge of elaborating the global evaluation plan for the 3 evaluation campaigns of the project. 47 participants joined in the TC-STAR evaluations, including both academic and industrial groups.

The aim of these evaluation campaigns was to measure the progress made during the project in:

- Automatic Speech Recognition (ASR),
- Spoken Language Translation (SLT),
- Text To Speech (TTS) processing,
- UIMA Integration of components (ASR+SLT, ASR+STL+TTS).

In order to be able to chain the ASR, SLT and TTS components, evaluation tasks were designed to use common sets of raw data and conditions. Three evaluation tasks, common to ASR and SLT, were selected:

- **European Parliament Plenary Sessions (EPPS):** the evaluation data consisted of audio recordings from the EPPS original channel(1) of the parliamentary debates, and of the official documents published by the European Community, containing post-edited transcriptions of the sessions, both in English and Spanish. The focus was exclusively on those Members of Parliament speaking in English and in Spanish, therefore, the speeches produced by the interpreters were not used. These resources were used to evaluate ASR in English and Spanish as well as SLT in the English-to-Spanish (En->Es) and Spanish-to-English (Es->En) directions.

- **CORTES Spanish Parliament Sessions:** since there are few Spanish speeches in the EPPS recordings, it was decided to use audio recordings of the Spanish Parliament (Congreso de Los Diputados). The data were used in addition to the EPPS Spanish data to evaluate ASR in Spanish and SLT from Spanish into English (Es->En).

(1) This channel includes speeches from Members of the Parliament in their original language.

- **Voice Of America:** The evaluation data consisted of audio recordings in Mandarin Chinese (Zh), from the broadcast news of the Mandarin "Voice of America" (VOA) radio station. Those data were used to evaluate speech recognition systems in Mandarin Chinese and translation from Mandarin into English (Zh->En).

With regard to ASR, the Word Error Rate (WER) or Character Error Rate has been widely reduced, as it can be seen in Figure 1, through the results obtained by some of the project participants.

Regarding SLT, systems participating in the campaigns were evaluated on 3 directions (Zh->En, En->Es and Es->En) and on three kinds of input (ASR output, verbatim and well formatted text). Automatic metrics such as BLEU, NIST, IBM, mWER, mPER, and WNM were computed for this purpose. Further to the

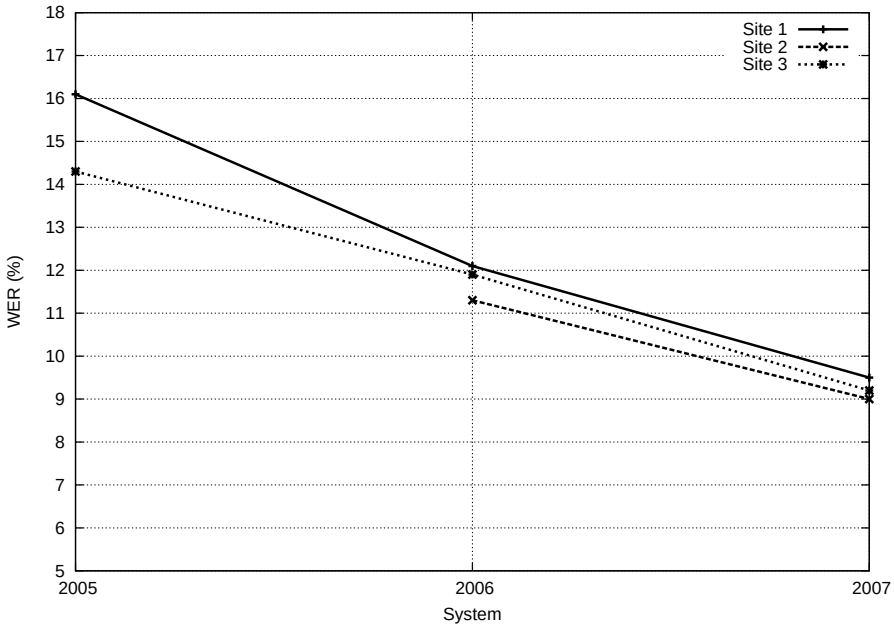


Figure 1: Progress on ASR for English

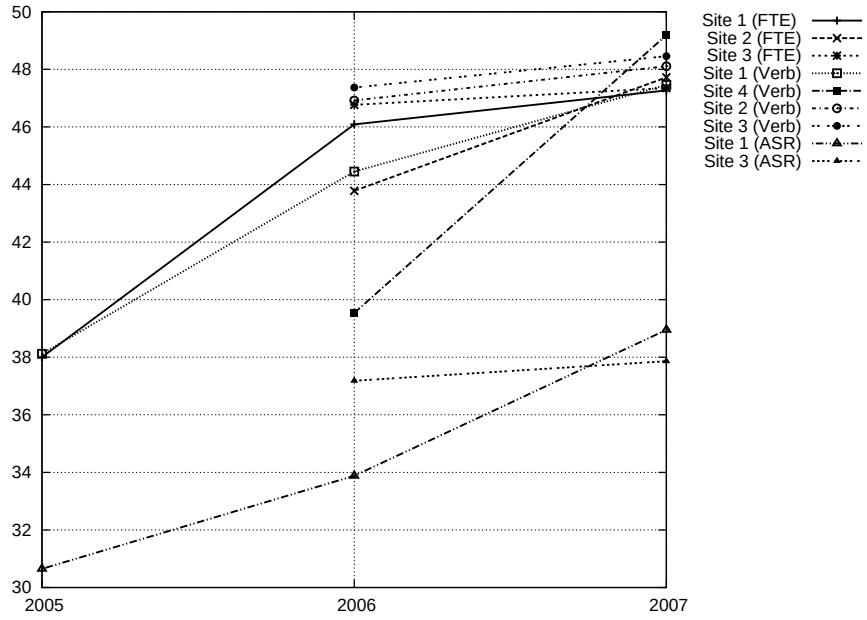


Figure 2: Spanish-to-English (EPPS&CORTES) improvement

Evaluation task	Language
Text processing	
1. Non Standard Word Normalization	En
2. End-of-sentence detection (En) or Words segmentation (Zh)	Zh, En
3. POS (Part-Of-Speech) Tagging	Zh, En
4. Grapheme-to-phoneme conversion	Zh, En
Prosody Generation	
5. Use of segmental information	En, Es
6. Rating of delexicalised utterances	Zh, En, Es
7. Choice of a delexicalised utterance	Zh, En, Es
Acoustic Synthesis	
8. Intelligibility test	Zh, En
9. Judgment test	Zh, En
Intra-lingual Voice Conversion (IVC)	
10. Comparison of speaker identities	Zh, En, Es
11. Evaluation of overall speech quality	Zh, En, Es
Crosslingual Voice Conversion (CVC)	
12. Comparison of speaker identities	En/Es, Es/En
13. Evaluation of overall speech quality	En/Es, Es/En
Expressive speech	
14. Judgment test	Es
15. Comparison test	Es
TTS Component	
16. judgement test	Zh, En, Es
17. Intelligibility test	Zh, En, Es

Table 1: TTS evaluation tasks

automatic metrics, important human evaluations were also carried out. Each segment was evaluated in relation to *adequacy* and *fluency* measures. For the evaluation of adequacy, the target segment was compared to a reference segment. For the evaluation of fluency, only the syntactic quality of the translation was evaluated. The evaluators graded all the segments firstly according to fluency, and then according to adequacy, so that both types of measures were done independently, but making sure that each evaluator did both for a certain number of segments. The correlation between metrics was also studied.

As for SLT, Figure 2 shows the important improvement of MT modules during the 3 years of the project.

The evaluation of speech synthesis was divided into 17 different tasks, involving 3 languages: English (En), Spanish (Es) and Mandarin Chinese (Zh). Table 1 lists the TTS evaluation tasks considered together with the languages involved.

Except for the text processing tasks, evaluations consisted in subjective tests. In addition to the evaluations of individual components, an end-to-end evaluation of the whole speech-to-speech translation process was also carried out. This end-to-end evaluation consisted in evaluating the full speech-to-speech system, by chaining one ASR system, one SLT system and one TTS system.

The end-to-end evaluation was carried out in one translation direction, English-to-Spanish, and using 2 measures:

- Adequacy test: comprehension test on potential users which allows measuring the intelligibility rate;
- Fluency test: judgment test with several questions related to fluency and also usability of the system.

At present, all TC-STAR packages are being made available to assess system performance and a number of copies have already been distributed.

Closely related to the TC-STAR project was also part of the work carried out within the ECESS (European Center of Excellence for Speech Synthesis) evaluation framework. ELDA has collaborated in the setting up of an evaluation framework regarding software modules and tools for TTS. To present, the evaluation of modules has focused on text processing, prosody and acoustic synthesis, and the evaluation of tools has concerned pitch extraction, voice activity detection and phonetic segmentation.

CHIL

Another major project worth mentioning here is **CHIL** ("Computers in the Human Interaction Loop", an FP6 IP). The implication of ELDA made it easy to capitalize on the work conducted within the project and to ensure that all data sets and evaluation toolkits are made widely and immediately available under very fair conditions. CHIL has addressed the largest number of technology components ever done before with the goal to develop computer assistants that attend to human activities, interactions, and intentions. CHIL's thirteen technological components were evaluated, which focused on Vision technologies (Face Detection, Visual Person Tracking, Visual Speaker Identification, Head Pose Estimation, Hand Tracking), on Sound and Speech technologies (Close-Talking Automatic Speech Recognition, Far-Field Automatic Speech Recognition, Acoustic Person Tracking, Acoustic Speaker Identification, Speech Activity Detection, Acoustic Scene Analysis) and on Contents Processing technologies (Automatic Summarization and Question Answering on Spoken Transcriptions, conducted in partnership with CLEF). The corresponding evaluation packages are being made available through ELRA's Catalogue.

A further international achievement resulting from this project, and of great importance, is the establishment of the open international evaluation workshop **CLEAR** - "CLassification of Events, Activities, and Relationships", in partnership with NIST and other players. So far, two CLEAR evaluation campaigns were conducted, partly with CHIL packages.

In CLEAR 2007, ELDA organized the evaluation of the following technologies:

- Vision technologies:

  - *Face Detection and Tracking*. The goal of the face tracking task is to detect the faces in each frame and track them throughout the given sequence.

  - *Visual Person Tracking*. The goal is to continuously and simultaneously track all attendees of an interactive seminar for the length of a sequence using all available cameras.

  - *Visual Speaker Identification*. The goal is to identify a closed-set of people based on visual data streams. Systems shall provide an identity estimate for each test segment.

  - *Head Pose Estimation*. The goal of this task is to estimate the head orientation of people from respective camera observations.

- Audio technologies:

  - *Acoustic Person Tracking*. The goal is to detect speech activity and to track the respective speaker in segments of non-overlapping speech using all available far-field microphones.

  - *Acoustic Speaker Identification*. The goal is to identify a closed-set of people based on acoustic data streams.

  - *Acoustic Event Detection*. The goal of this task is to detect and recognize a closed set of pre-defined acoustic events.

- Multimodal technologies:

  - *Multimodal Person Tracking*. The goal is to detect speaker turns and to audio-visually track the last known speaker, even through periods of silence or noise, using all available sensors, cameras and microphones.

  - *Multimodal Person Identification*. The goal is to identify a closed-set of people based on audio-visual data streams.

CLEF

Another major area being tackled by most of the HLT key players is the Multilingual/Cross-Lingual Information Access and Retrieval. Through some partial European funding, CLEF (Cross-Language Evaluation Forum) was launched in 2000 with the aim to develop an infrastructure for the evaluation, testing and tuning of information retrieval systems operating on European languages in both monolingual and cross-language contexts and, beyond this, to experiment the setting up of a European HLT evaluation infrastructure. The project managed to create test suites of reusable data which are part of the ELRA evaluation catalogue and which are extensively employed by system developers for benchmarking purposes. The exploitation of the methodologies implemented by CLEF for the testing and tuning of information retrieval systems is now part of ELDA's assets and it allows conducting the evaluation of commercial products and applications with a strong and reliable technical and scientific background. More than 11 copies of the CLEF packages have been distributed so far.

Within CLEF 2008, ELDA is co-organizing a cross-language adaptive filtering evaluation campaign called INFILE.

INFILE extends the last filtering track of TREC-2002 in the following ways:

- INFILE is crosslingual (English, French and Arabic): a corpus of 100,000 comparable newswire stories from Agence France Presse (AFP) is used for evaluation in each of the languages considered.
- Evaluation will be performed using an automatic interrogation of test systems with a simulated user feedback. Each system will be able to use the feedback at any time to increase performance.

Technolangue/Evalda Programme

Another important initiative, funded by a national agency, is the French programme "Technolangue/Evalda": the Evalda projects that ELDA has coordinated consisted of 8 evaluation campaigns with a focus on the spoken and written language technologies for the French language: ARCADE II (evaluation of bilingual corpora alignment systems), CESART (evaluation of terminology extraction systems), CESTA (evaluation of machine translation systems), EASY (evaluation of parsers), ESTER (evaluation of broadcast news automatic transcribing systems), EQUER (evaluation of question answering systems), EVASY (evaluation of speech synthesis systems), and MEDIA (evaluation of in and out-of context dialog systems). As planned and achieved within the other evaluation projects, all Evalda evaluation resources have been packaged and made available and more than 15 copies have been distributed so far.

Web-based Evaluation Services

ELDA has recently invested a major effort on developing web-based services for HLT evaluations. Examples of automated evaluation platforms deployed as a web server are currently very rare and often underestimated. However, the benefits offered by such services are plentiful. Time gain is one of such benefits, together with effort savings, faster system improvement, and the very idea behind providing a common paradigm of evaluation for the community.

The first web-services were developed for evaluating ASR and MT modules within the TC-STAR project. Participants could register their systems into the web service, download the audio recordings, upload the automatic transcriptions and obtain their results in terms of word error rate immediately, which made the whole process fast, efficient, and comparable for different technology developers. The same architecture was available for the evaluation of Machine Translation. These services were developed within the UIMA framework.

Moreover, ELDA deployed another web evaluation service during an evaluation campaign of French parsers. This campaign took place within the earlier-mentioned "Technolangue/Evalda" programme and it certainly proved the interest of evaluation web services by allowing participants to concentrate on the development of their systems and by making feasible the scoring of hundreds of submissions. We are currently planning to generalise such service to other NLP domains.

Further to its participation in such projects, ELDA has also run a number of initiatives, such as discussion and brainstorming events. One such example is ELRA's 10th Anniversary Workshop on Evaluation, celebrated in Malta in December 2005. The discussions initiated at this occasion have been taken further with the Evaluation Workshop celebrated during the MT Summit 2007 (Automatic Procedures in MT Evaluation), and a coming ELRA Evaluation Workshop (Looking into the Future of Evaluation:

when automatic metrics meet task-based and performance-based approaches) to be celebrated jointly with the LREC 2008 Conference, this coming May-June 2008.

Conclusion

As stated above, ELRA has been entrusted with a crucial mission: to ensure that Language Resources needed by Language Engineering players are made available when they already exist or to produce them in a cost-effective frame. It is of paramount importance that regional organizations emerge and co-operate between themselves with respect to the issues described herein. The main common task would be to achieve, all together, a better streamlining of efforts in the development of new Language Resources that are of interest to "local" and "global" players. This role should be extended to Evaluation in particular in geographical areas that do not have a dedicated organization.

At the same time, the paradigm of evaluation should be reconsidered by the funding agencies and funded as a major part of their investments, as it allows both to measure if the money they have invested in technology development has led to significant progress and to identify areas where the technology needs further improvement.

Evaluation also allows application developers/integrators and end-users to understand where the technology is and how it can help them and provide them with new solutions to the problems they face.

In its involvement in HLT evaluation, ELDA already covers a wide range of technologies, such as:

- Audio Technologies:
  - Automatic Speech Recognition (ASR),
  - Spoken Language Translation (SLT),
  - Text-To-Speech (TTS) processing (covering tasks like: Text Processing, Prosody Generation, Acoustic Synthesis, Intra-lingual Voice Conversion, Crosslingual Voice Conversion, Expressive Speech and TTS Components),
  - Acoustic Person Tracking,
  - Acoustic Speaker Identification,
  - Acoustic Event Detection,
  - Acoustic Scene Analysis,
  - In- and Out-of-Context Dialog Systems.
- Vision Technologies:
  - Face Detection and Tracking,
  - Visual Person Tracking,
  - Visual Speaker Identification,
  - Head Pose Estimation,
  - Hand Tracking.
- Multimodal Technologies:
  - Multimodal Person Tracking,
  - Multimodal Person Identification.
- Text Processing Technologies:
  - Question Answering,
  - Information Retrieval/Filtering,
  - Automatic Summarization,
  - Machine Translation,
  - PoS Tagging & Parsing,
  - Corpora Alignment,
  - Terminology Extraction.

Finally, ELRA also initiated the HLT Evaluation portal that is designed to be an online information resource about HLT evaluation and related topics of interest to the HLT community at large.

Links

1) TC-STAR project:
http://www.tc-star.org

2) CHIL project:
http://chil.server.de/servlet/is/101

3) ECESS:
http://www.ecess.eu

4) CLEAR:
http://www.clear-evaluation.org

5) CLEF:
http://www.clef-campaign.org

6) Technolangue/Evalda programme:
http://www.technolangue.net/rubrique.php3?id_rubrique=24

7) ELRA HLT Evaluation Portal:
http://www.hlt-evaluation.org

NEW RESOURCES

Monolingual Lexicons from the general domain

ELRA-L0076 Polderland Dutch Lexicon of Abbreviations and Acronyms

The lexicon contains 2,180 Dutch abbreviations and acronyms. It complies with the official Dutch Spelling (2005/6). Each entry consists of an ID, word form, lemma and part of speech.
The lexicon is delivered in human-readable UNIX ANSI-format, with just a linefeed at the end of every line instead of a carriage return. The character set used is ISO Latin-1 (ISO/IEC 8859-1). All fields are separated by horizontal tabs, with no empty fields.

	ELRA members	Non-members
For research use	200 Euro	260 Euro
For commercial use	260 Euro	350 Euro

ELRA-L0077 Polderland Dutch General Lexicon

The lexicon contains 400,463 Dutch words, comprising 236,369 nouns, 90,882 adjectives, 69,744 verbs, 2,120 adverbs, and 1,348 items from other categories (pronouns, determiners, articles, adpositions, conjunctions, numerals, etc.). It complies with the official Dutch Spelling (2005/6). The lexicon contains an ID, word form, lemma and part of speech.
The lexicon is delivered in human-readable UNIX ANSI-format, with just a linefeed at the end of every line instead of a carriage return. The character set used is ISO Latin-1 (ISO/IEC 8859-1). All fields are separated by horizontal tabs, with no empty fields.

	ELRA members	Non-members
For research use	38,000 Euro	48,000 Euro
For commercial use	48,000 Euro	60,000 Euro

ELRA-L0078 Polderland Dutch Lexicon of Names

The lexicon contains 24,247 Dutch proper names. Various sorts of proper names are included, such as first names, last names, geographical names etc. Each entry contains an ID, word form, lemma, part of speech and proper name type.

The lexicon is delivered in human-readable UNIX ANSI-format, with just a linefeed at the end of every line instead of a carriage return. The character set used is ISO Latin-1 (ISO/IEC 8859-1). All fields are separated by horizontal tabs, with no empty fields.

	ELRA members	Non-members
For research use	2,300 Euro	2,900 Euro
For commercial use	2,900 Euro	3,800 Euro

ELRA-L0079 Polderland Dutch Lexicon of Business Terminology

The lexicon contains 15,987 Dutch words from the business domain, comprising 13,774 nouns, 1,267 adjectives, 895 verbs, 9 adverbs, and 42 items from other categories. The lexicon complies with the official Dutch Spelling (2005). Each entry contains an ID, word form and part of speech.

The lexicon is delivered in human-readable UNIX ANSI-format, with just a linefeed at the end of every line instead of a carriage return. The character set used is ISO Latin-1 (ISO/IEC 8859-1). All fields are separated by horizontal tabs, with no empty fields.

	ELRA members	Non-members
For research use	1,500 Euro	1,900 Euro
For commercial use	1,900 Euro	2,400 Euro

ELRA-L0080 Polderland Dutch Lexicon of Legal Terminology

The lexicon contains 6,207 Dutch words from the legal domain, comprising 4,781 nouns, 810 adjectives, 573 verbs, 12 adverbs and 31 items from other categories. It complies with the official Dutch Spelling (2005/6). Each entry contains an ID, word form and part of speech.

The lexicon is delivered in human-readable UNIX ANSI-format, with just a linefeed at the end of every line instead of a carriage return. The character set used is ISO Latin-1 (ISO/IEC 8859-1). All fields are separated by horizontal tabs, with no empty fields.

	ELRA members	Non-members
For research use	600 Euro	750 Euro
For commercial use	750 Euro	950 Euro

ELRA-L0081 Polderland Dutch Lexicon of Medical Terminology

The lexicon contains 17,115 Dutch words from the medical domain, comprising 12,638 nouns, 3,107 adjectives, 1,273 verbs, 11 adverbs and 86 items from other categories. It complies with the official Dutch Spelling (2005/6). Each entry contains an ID, word form and part of speech.

The lexicon is delivered in human-readable UNIX ANSI-format, with just a linefeed at the end of every line instead of a carriage return. The character set used is ISO Latin-1 (ISO/IEC 8859-1). All fields are separated by horizontal tabs, with no empty fields.

	ELRA members	Non-members
For research use	1,600 Euro	2,100 Euro
For commercial use	2,100 Euro	2,600 Euro

ELRA-L0082 Polderland Dutch Lexicon of Social Terminology

The lexicon contains 12,551 Dutch words from the social domain, comprising 9,984 nouns, 1,306 adjectives, 1,161 verbs, 56 adverbs and 44 items from other categories. It complies with the official Dutch Spelling (2005/6). Each entry contains an ID, word form and part of speech.

The lexicon is delivered in human-readable UNIX ANSI-format, with just a linefeed at the end of every line instead of a carriage return. The character set used is ISO Latin-1 (ISO/IEC 8859-1). All fields are separated by horizontal tabs, with no empty fields.

	ELRA members	Non-members
For research use	1,200 Euro	1,500 Euro
For commercial use	1,500 Euro	1,900 Euro

ELRA-L0083 Polderland Dutch Lexicon of Technical Terminology

The lexicon contains 9,940 Dutch words from the technical/scientific domain, comprising 8,832 nouns, 950 adjectives, 111 verbs, 2 adverbs and 45 items from other categories. It complies with the official Dutch Spelling (2005/6). Each entry contains an ID, word form and part of speech.

The lexicon is delivered in human-readable UNIX ANSI-format, with just a linefeed at the end of every line instead of a carriage return. The character set used is ISO Latin-1 (ISO/IEC 8859-1). All fields are separated by horizontal tabs, with no empty fields.

	ELRA members	Non-members
For research use	900 Euro	1,200 Euro
For commercial use	1,200 Euro	1,500 Euro

ELRA-L0084 Macedonian Morphological Lexicon (MACPLEX)

MACPLEX comprises two dictionaries: a dictionary of lemmas (over 80,000 entries) and a dictionary of word forms (over 1,300,000 entries). Morphological information (PoS, gender, case, definiteness, number for nouns, tense, person, etc. for verbs) is available for each entry. Out of the more than 1,300,000 word forms, there are 345,350 nouns, 467,744 adjectives, 500,220 verbs and 19,472 adverbs. The remaining entries correspond to pronouns, adpositions, conjunctions and numerals. The lexicon is available in Unicode.

	ELRA members	Non-members
For research use by academic organisations	2,000 Euro	2,500 Euro
For research use by commercial organisations	3,000 Euro	4,000 Euro
For commercial use	8,000 Euro	10,000 Euro

Phonetic lexicons from the LC-STAR European-funded project (Lexica and Corpora for Speech-to-Speech Translation Components)

ELRA-S0245 LC-STAR German Phonetic Lexicon

The LC-STAR German Phonetic lexicon was created within the scope of the LC-STAR project (IST 2001-32216) which was sponsored by the European Commission.
The lexicon comprises 102,169 entries, distributed over three categories:
- a set of 55,507 common word entries. This set is extracted from a corpus of more than 15 million words distributed over 6 different domains (sports/games, news, finance, culture/entertainment, consumer information, personal communications). This was done with the aim of reaching a target for each domain of at least 95% self coverage. In addition to extracting word lists from the corpus, a list of closed set (function) word classes are included in the final word list.
- a set of 46,662 proper names (including person names, family names, cities, streets, companies and brand names) divided into 3 domains. Multiple word names such as New_York are kept together in all three domains, and they count as one entry. The 3 domains consist of first and last names (20,864 different entries), place names (22,212 different entries), and organisations (5,523 different entries).
- and a list of 6,763 special application words translated from English terms defined by the LC-STAR consortium. This list contains: numbers, letters, abbreviations and specific vocabulary for applications controlled by voice (information retrieval, controlling of consumer devices, etc.).

The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA. It is stored on 1 CD.

	ELRA members	Non-members
For research use	21,250 Euro	27,625 Euro
For commercial use	28,000 Euro	36,400 Euro

ELRA-S0246 LC-STAR German Phonetic Lexicon in the Touristic Domain

The LC-STAR German Phonetic lexicon was created within the scope of the LC-STAR project (IST 2001-32216) which was sponsored by the European Commission.
The lexicon comprises 8,782 entries from the following categories: nouns, adjectives and verbs.
For each entry the following information is provided:
- orthographic form (spelling)
- part-of-speech (POS) with grammatical attributes and lemma
- phonemic transcription

The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA. It is stored on 1 CD.

	ELRA members	Non-members
For research use	2,000 Euro	2,600 Euro
For commercial use	3,000 Euro	3,900 Euro

ELRA-S0247 LC-STAR Standard Arabic Phonetic Lexicon

The LC-STAR Standard Arabic Phonetic lexicon was created within the scope of the LC-STAR project (IST 2001-32216) which was sponsored by the European Commission.
The lexicon comprises 110,271 entries, distributed over three categories:
- a set of 52,981 common word entries. This set is extracted from a corpus of more than 13 million words distributed over 6 different domains (sports/games, news, finance, culture/entertainment, consumer information, personal communications). This was done with the aim of reaching a target for each domain of at least 95% self coverage. In addition to extracting word lists from the corpus, a list of closed set (function) word classes are included in the final word list.
- a set of 50,135 proper names (including person names, family names, cities, streets, companies and brand names) divided into 3 domains. Multiple word names such as New_York are kept together in all three domains, and they count as one entry. The 3 domains consist of first and last names (9,738 different entries), place names (22,998 different entries), and organisations (17,309 different entries).
- and a list of 7,155 special application words translated from English terms defined by the LC-STAR consortium. This list contains: numbers, letters, abbreviations and specific vocabulary for applications controlled by voice (information retrieval, controlling of consumer devices, etc.).

The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA. It is stored on 1 CD.

	ELRA members	Non-members
For research use	21,250 Euro	27,625 Euro
For commercial use	28,000 Euro	36,400 Euro

ELRA-S0248 LC-STAR English-German Bilingual Aligned Phrasal Lexicon

The LC-STAR English-German Bilingual Aligned Phrasal lexicon was created within the scope of the LC-STAR project (IST 2001-32216) which was sponsored by the European Commission. It was designed for SST (Speech-to-Speech Translation).
The lexicon comprises 10,733 phrases from the tourist domain. It is based on a list of short sentences obtained by translation from a US-English 10,518 phrase corpus. The total number of unique separate words is 8,782.
The lexicon contains the following information: US-English phrase (orthography), its translation into German (orthography), and for each token in German a phrase provides the following: orthography of a word, part of speech, lemma, whether the phrase is idiomatic or not, if a word is a foreign word. In this lexicon, foreign words were only tagged if they were written with foreign orthography (e.g. English characters). The lexicon is provided in XML format. It is stored on 1 CD.

	ELRA members	Non-members
For research use	3,750 Euro	4,875 Euro
For commercial use	5,500 Euro	7,150 Euro

ELRA-S0255 LC-STAR Finnish Phonetic Lexicon

The LC-STAR Finnish Phonetic lexicon was created within the scope of the LC-STAR project (IST 2001-32216) which was sponsored by the European Commission.
The lexicon comprises 189,409 entries, distributed over three categories:
- a set of 144,233 common word entries. This set is extracted from a corpus of more than 18 million words distributed over 6 different domains (sports/games, news, finance, culture/entertainment, consumer information, personal communications). This was done with the aim of reaching a target for each domain of at least 95% self coverage. In addition to extracting word lists from the corpus, a list of closed set (function) word classes are included in the final word list.
- a set of 45,176 proper names (including person names, family names, cities, streets, companies and brand names) divided into 3 domains. Multiple word names such as New_York are kept together in all three domains, and they count as one entry. The 3 domains consist of first and last names (21,903 different entries), place names (13,168 different entries), and organisations (10,541 different entries).
- and a list of 13,068 special application words translated from English terms defined by the LC-STAR consortium. This list contains: numbers, letters, abbreviations and specific vocabulary for applications controlled by voice (information retrieval, controlling of consumer devices, etc.).
The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA. It is stored on 1 CD.

	ELRA members	Non-members
For research use	15,000 Euro	22,000 Euro
For commercial use	23,000 Euro	30,000 Euro

ELRA-S0256 LC-STAR Mandarin Chinese Phonetic Lexicon

The LC-STAR Mandarin Chinese Phonetic lexicon was created within the scope of the LC-STAR project (IST 2001-32216) which was sponsored by the European Commission.
The lexicon comprises 104,368 entries, distributed over three categories:
- a set of 38,098 common word entries. This set is extracted from a corpus of more than 20 million words distributed over 6 different domains (sports/games, news, finance, culture/entertainment, consumer information, personal communications). This was done with the aim of reaching a target for each domain of at least 95% self coverage. In addition to extracting word lists from the corpus, a list of closed set (function) word classes are included in the final word list.
- a set of 57,528 proper names (including person names, family names, cities, streets, companies and brand names) divided into 3 domains. Multiple word names such as New_York are kept together in all three domains, and they count as one entry. The 3 domains consist of first and last names (22,141 different entries), place names (19,872 different entries), and organisations (15,600 different entries).
- and a list of 7,522 special application words translated from English terms defined by the LC-STAR consortium. This list contains: numbers, letters, abbreviations and specific vocabulary for applications controlled by voice (information retrieval, controlling of consumer devices, etc.).
The lexicon is provided in XML format and includes phonetic transcriptions in SAMPA. It is stored on 1 CD.

	ELRA members	Non-members
For research use	27,000 Euro	38,000 Euro
For commercial use	40,000 Euro	50,000 Euro

ELRA-S0257 LC-STAR English-Finnish Bilingual Aligned Phrasal Lexicon

The LC-STAR English-Finnish Bilingual Aligned Phrasal lexicon was created within the scope of the LC-STAR project (IST 2001-32216) which was sponsored by the European Commission. It was designed for SST (Speech-to-Speech Translation).
The lexicon comprises 10,520 phrases from the tourist domain. It is based on a list of short sentences obtained by translation from a US-English 10,518 phrase corpus. The total number of unique separate words is 28,568.
The lexicon contains the following information: US-English phrase (orthography), its translation into Finnish (orthography), and for each token in Finnish a phrase provides the following: orthography of a word, part of speech, lemma, whether the phrase is idiomatic or not, if a word is a foreign word. In this lexicon, foreign words were only tagged if they were written with foreign orthography (e.g. English characters). The lexicon is provided in XML format. It is stored on 1 CD.

	ELRA members	Non-members
For research use	3,750 Euro	4,875 Euro
For commercial use	5,500 Euro	7,150 Euro

Speech Microphone resources from the Speecon European-funded project (for the development of voice controlled consumer applications)

ELRA-S0244 Japanese Speecon Database

The Japanese Speecon database comprises the recordings of 556 adult Japanese speakers and 51 child Japanese speakers who uttered respectively over 290 items and 210 items (read and spontaneous).

ELRA-S0265 Dutch from Belgium Speecon Database

The Dutch from Belgium Speecon database comprises the recordings of 550 adult speakers and 50 child speakers who uttered respectively over 290 items and 210 items (read and spontaneous).

ELRA-S0266 Dutch from the Netherlands Speecon Database

The Dutch from the Netherlands Speecon database comprises the recordings of 550 adult speakers and 50 child speakers who uttered respectively over 290 items and 210 items (read and spontaneous).

ELRA-S0267 Danish Speecon Database

The Danish Speecon database comprises the recordings of 550 adult speakers and 50 child speakers who uttered respectively over 290 items and 210 items (read and spontaneous).

Prices available upon request. Please contact us.

Speech Microphone resources from the TC-STAR European-funded project (for the development of ASR applications)

ELRA-S0249 TC-STAR English Training Corpora for ASR: Transcriptions of EPPS Speech

TC-STAR is a European integrated project focusing on all core technologies for Speech-to-Speech Translation (SST): Automatic Speech Recognition (ASR), Spoken Language Translation (SLT), and Text to Speech Synthesis (TTS).
This corpus consists of transcriptions from 92 hours of EPPS (European Parliament Plenary Sessions) speeches held or interpreted in European English (a mixture of native and non-native English). The recordings (not included in the present package) were obtained from Europe by Satellite (http://europa.eu.it/comm/ebs) from May 2004 until May 2006. The corpus consists of 63 transcriptions files. The transcription files are stored in Transcriber XML file format.
The speech databases made within the TC-STAR project were validated by SPEX, in the Netherlands, to assess their compliance with the TC-STAR format and content specifications.

For corresponding recordings, see ELRA-S0251.

	ELRA members	Non-members
For research use	4,400 Euro	5,750 Euro
For commercial use	6,250 Euro	8,200 Euro

ELRA-S0250 TC-STAR English-Spanish Training Corpora for Machine Translation: Aligned Final Text Editions of EPPS

TC-STAR is a European integrated project focusing on all core technologies for Speech-to-Speech Translation (SST): Automatic Speech Recognition (ASR), Spoken Language Translation (SLT), and Text to Speech Synthesis (TTS).
This corpus consists of respectively 34 million (English) and 38 million (Spanish) running words of bilingual sentence segmented and aligned texts in English and Spanish obtained from the Final Text Editions provided by the European Parliament (http://www.europarl.europa.eu) from April 1996 to Sept. 2004, Dec. 2004 to May 2005, and Dec. 2005 to May 2006. The data is accompanied by tools for further preprocessing.

	ELRA members	Non-members
For research use	3,000 Euro	3,925 Euro
For commercial use	4,250 Euro	5,600 Euro

ELRA-S0251 TC-STAR English Training Corpora for ASR: Recordings of EPPS Speech

TC-STAR is a European integrated project focusing on all core technologies for Speech-to-Speech Translation (SST): Automatic Speech Recognition (ASR), Spoken Language Translation (SLT), and Text to Speech Synthesis (TTS).
This corpus consists of the recordings of around 290 hours from EPPS (European Parliament Plenary Sessions) speeches held or interpreted in European English (a mixture of native and non-native English), 92 hours of which were annotated (transcribed) (the transcriptions are not included in the present package). These recordings were obtained from Europe by Satellite (http://europa.eu.it/comm/ebs) from May 2004 until May 2006.
The speech signals were submitted by EbS via internet in Real Media format and via satellite in MPEG1-layer2 format. The signals were decoded, resampled and are stored in WAVE RIFF (Resource Interchange File Format). Each file contains a single channel with 16-bit resolution at a sample rate of 16kHz.
The speech databases made within the TC-STAR project were validated by SPEX, in the Netherlands, to assess their compliance with the TC-STAR format and content specifications.

For corresponding transcriptions, see ELRA-S0249.

	ELRA members	Non-members
For research use	400 Euro	520 Euro
For commercial use	600 Euro	800 Euro

ELRA-S0252 TC-STAR Spanish Training Corpora for ASR: Recordings of EPPS Speech

TC-STAR is a European integrated project focusing on all core technologies for Speech-to-Speech Translation (SST): Automatic Speech Recognition (ASR), Spoken Language Translation (SLT), and Text to Speech Synthesis (TTS).
This corpus consists of the recordings of around 283 hours from EPPS (European Parliament Plenary Sessions) speeches held or interpreted in European Spanish (a mixture of native and non-native Spanish), 62 hours of which were annotated (transcribed) within the project (the transcriptions are not provided in the present package but will be made available soon). These recordings were obtained from Europe by Satellite (http://europa.eu.it/comm/ebs) from May 2004 until May 2006.
The speech signals were submitted by EbS via internet in Real Media format and via satellite in MPEG1-layer2 format. The signals were decoded, resampled and are stored in WAVE RIFF (Resource Interchange File Format). Each file contains a single channel with 16-bit resolution at a sample rate of 16kHz.

	ELRA members	Non-members
For research use	400 Euro	520 Euro
For commercial use	600 Euro	800 Euro

ELRA-S0253 TC-STAR English Test Corpora for ASR

TC-STAR is a European integrated project focusing on all core technologies for Speech-to-Speech Translation (SST): Automatic Speech Recognition (ASR), Spoken Language Translation (SLT), and Text to Speech Synthesis (TTS).
This corpus consists of 70 hours of recordings of EPPS (European Parliament Plenary Sessions) speeches held or interpreted in European English and other European languages. From this corpus, 16 hours of English speeches (native or non native) were annotated (transcribed). Transcriptions are included in the present package. The data comprises the test (development and evaluation) data for the TC-STAR project in the years 2005, 2006, and 2007. The recordings were obtained from Europe by Satellite (http://europa.eu.it/comm/ebs) from Oct. until Nov. 2004, June to Nov. 2005, and June until July 2006. The transcription files are stored in Transcriber XML file format.
The speech signals were submitted by EbS via internet in Real Media format and via satellite in MPEG1-layer2 format. The signals were decoded, resampled and are stored in WAVE RIFF (Resource Interchange File Format). Each file contains a single channel with 16-bit resolution at a sample rate of 16kHz.
The speech databases made within the TC-STAR project were validated by SPEX, in the Netherlands, to assess their compliance with the TC-STAR format and content specifications.

	ELRA members	Non-members
For research use	1,500 Euro	2,250 Euro
For commercial use	4,500 Euro	6,750 Euro

ELRA-S0254 TC-STAR Spanish Test Corpora for ASR

TC-STAR is a European integrated project focusing on all core technologies for Speech-to-Speech Translation (SST): Automatic Speech Recognition (ASR), Spoken Language Translation (SLT), and Text to Speech Synthesis (TTS).
This corpus consists of 174 hours of recordings of EPPS (European Parliament Plenary Sessions) speeches held or interpreted in European Spanish and other European languages. From this corpus, 16 hours of Spanish speeches were annotated (transcribed). Transcriptions are included in the present package. The data comprises the test (development and evaluation) data for the TC-STAR project in the years 2005, 2006, and 2007. The recordings were obtained from Europe by Satellite (http://europa.eu.it/comm/ebs) from Oct. until Nov. 2004, June to Nov. 2005, and June until Sept. 2006. The transcription files are stored in Transcriber XML file format.
The speech signals were submitted by EbS via internet in Real Media format and via satellite in MPEG1-layer2 format. The signals were decoded, resampled and are stored in WAVE RIFF (Resource Interchange File Format). Each file contains a single channel with 16-bit resolution at a sample rate of 16kHz.
The speech databases made within the TC-STAR project were validated by SPEX, in the Netherlands, to assess their compliance with the TC-STAR format and content specifications.

	ELRA members	Non-members
For research use	1,500 Euro	2,250 Euro
For commercial use	4,500 Euro	6,750 Euro

Speech Telephone Database from the Orientel European-funded project (Multilingual Access to Interactive Communication Services for the Mediterranean & the Middle East)

ELRA-S0258 Orientel United Arab Emirates MCA (Modern Colloquial Arabic)

This speech database contains the recordings of 750 Arabic speakers recorded over the United Arab Emirates' fixed and mobile telephone network. Each speaker uttered around 48 read and spontaneous items.

	ELRA members	Non-members
For research use	26,600 Euro	33,250 Euro
For commercial use	28,000 Euro	35,000 Euro

Special prices for a combined purchase of S0258, S0259 and S0260.

ELRA-S0259 Orientel United Arab Emirates MSA (Modern Standard Arabic)

This speech database contains the recordings of 500 Arabic speakers recorded over the United Arab Emirates' fixed and mobile telephone network. Each speaker uttered around 49 read and spontaneous items.

	ELRA members	Non-members
For research use	17,100 Euro	21,375 Euro
For commercial use	18,000 Euro	22,500 Euro

Special prices for a combined purchase of S0258, S0259 and S0260.

ELRA-S0260 Orientel English as spoken in the United Arab Emirates

This speech database contains the recordings of 535 speakers of English recorded over the United Arab Emirates' fixed and mobile telephone network. Each speaker uttered around 51 read and spontaneous items.

	ELRA members	Non-members
For research use	17,100 Euro	21,375 Euro
For commercial use	18,000 Euro	22,500 Euro

Special prices for a combined purchase of S0258, S0259 and S0260.

Speech Telephone from the SpeechDat(E) European-funded project (Eastern European Speech Databases for Creation of Voice Driven Teleservices)

ELRA-S0261 Hungarian SpeechDat(E) Database

This speech database contains the recordings of 1,000 Hungarian speakers recorded over the Hungarian fixed telephone network. Each speaker uttered around 50 read and spontaneous items.

	ELRA members	Non-members
For research use	15,200 Euro	21,375 Euro
For commercial use	16,000 Euro	22,500 Euro

Speech Telephone based on the SpeechDat project conventions

ELRA-S0243 SpeechDat Catalan FDB Database

The SpeechDat Catalan FDB database contains the recordings of 1,005 Catalan speakers (474 males, 531 females) recorded over the Spanish fixed telephone network. The database is partitioned into 4 CD-ROMs, in ISO 9660 format.

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 an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items: 3 application words; 1 sequence of 10 isolated digits; 4 connected digits (prompt sheet number -6 digits, telephone number -9/11 digits, credit card number -14/16 digits, PIN code -6 digits); 3 dates (spontaneous date e.g. birthday, prompted date, relative and general date expression); 1 word spotting phrase using embedded application words; 1 isolated digit; 3 spelled words (1 surname, 1 directory assistance city name, 1 real/artificial name for coverage); 1 currency money amount; 1 natural number; 5 directory assistance names (1 spontaneous, e.g. own surname, 1 city of birth/growing up, 1 most frequent city out of a set of 500, 1 most frequent company/agency out of a set of 500, 1 "forename surname" out of a set of 150); 2 yes/no questions (1 predominantly "yes" question, 1 predominantly "no" question, including fuzzy questions); 9 phonetically rich sentences; 2 time phrases (1 spontaneous time of day, 1 word style time phrase); 4 phonetically rich words

The following age distribution has been obtained: 13 speakers are under 16, 473 are between 16 and 30, 286 are between 31 and 45, 192 are between 46 and 60, and 41 speakers are over 60.

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

	ELRA members	Non-members
For research use	9,000 Euro	20,000 Euro
For commercial use	14,000 Euro	25,000 Euro

Speech Telephone Database from the SALA European-funded project (SpeechDat Across Latin America)

ELRA-S0242 SALA II US English Database

The SALA II US English database comprises 3,065 US English speakers (1,515 males, 1,550 females, including some speakers with Hispanic accents) recorded over the United States mobile telephone network.

	ELRA members	Non-members
For research use	41,250 Euro	45,000 Euro
For commercial use	45,000 Euro	56,250 Euro

ELRA-S0262 SALA II Portuguese from Brazil Database

The SALA II Portuguese from Brazil database comprises 1,000 Brazilian speakers recorded over the Brazilian mobile telephone network.

	ELRA members	Non-members
For research use	42,750 Euro	48,450 Euro
For commercial use	45,000 Euro	51,000 Euro

ELRA-S0263 SALA II Spanish from Colombia Database

The SALA II Spanish from Colombia database comprises 1,000 Colombian speakers recorded over the Colombian mobile telephone network.

	ELRA members	Non-members
For research use	42,750 Euro	48,450 Euro
For commercial use	45,000 Euro	51,000 Euro

ELRA-S0264 SALA II US Spanish West Database

The SALA II US Spanish West database comprises 1,000 Spanish speakers recorded over the American mobile telephone network.

	ELRA members	Non-members
For research use	42,750 Euro	48,450 Euro
For commercial use	45,000 Euro	51,000 Euro

Evaluation Packages

AURORA-CD0005 AURORA-5

The Aurora project was originally set up to establish a worldwide standard for the feature extraction software which forms the core of the front-end of a DSR (Distributed Speech Recognition) system.

The AURORA-5 database has been mainly developed to investigate the influence on the performance of automatic speech recognition for a hands-free speech input in noisy room environments. Furthermore two test conditions are included to study the influence of transmitting the speech in a mobile communication system.

The earlier three Aurora experiments had a focus on additive noise and the influence of some telephone frequency characteristics. Aurora-5 tries to cover all effects as they occur in realistic application scenarios. The focus was put on two scenarios. The first one is the hands-free speech input in the noisy car environment with the intention of controlling either devices in the car itself or retrieving information from a remote speech server over the telephone. The second one covers the hands-free speech input in a type of office or in a type of living room to control e.g. a telephone device or some audio/video equipment.

The AURORA-5 database contains the following data:

- Artificially distorted versions of the recordings from adult speakers in the TI-Digits speech database downsampled at a sampling frequency of 8000 Hz. The distortions consist of:
  - additive background noise,
  - the simulation of a hands-free speech input in rooms,
  - the simulation of transmitting speech over cellular telephone networks.
- A subset of recordings from the meeting recorder project at the International Computer Science Institute. The recordings contain sequences of digits uttered by different speakers in hands-free mode in a meeting room.
- A set of scripts for running recognition experiments on the above mentioned speech data. The experiments are based on the usage of the freely available software package HTK where HTK is not part of this resource.

	ELRA members	Non-members
For research use	250 Euro*	250 Euro*

* The TI digits are included in this package and must have been obtained from the LDC (ref. LDC93S10), priorily to any order of AURORA-5.

TC-STAR Evaluation Packages

The Evaluation Packages below include the material used for the TC-STAR 2007 Automatic Speech Recognition (ASR) and Spoken Language Translation (SLT) third evaluation campaign, as well as the material used for the TC-STAR 2006 and 2007 End-to-End task. They include resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself. The following TC-STAR Evaluation Packages are available:

- **ELRA-E0025** **TC-STAR 2007 Evaluation Package - ASR English**
- **ELRA-E0026-01** **TC-STAR 2007 Evaluation Package - ASR Spanish - CORTES**
- **ELRA-E0026-02** **TC-STAR 2007 Evaluation Package - ASR Spanish - EPPS**
- **ELRA-E0027** **TC-STAR 2007 Evaluation Package - ASR Mandarin Chinese**
- **ELRA-E0028** **TC-STAR 2007 Evaluation Package - SLT English-to-Spanish**
- **ELRA-E0029-01** **TC-STAR 2007 Evaluation Package - SLT Spanish-to-English - CORTES**
- **ELRA-E0029-02** **TC-STAR 2007 Evaluation Package - SLT Spanish-to-English - EPPS**
- **ELRA-E0030** **TC-STAR 2007 Evaluation Package - SLT Chinese-to-English**
- **ELRA-E0031** **TC-STAR 2006 Evaluation Package - End-to-End**
- **ELRA-E0032** **TC-STAR 2007 Evaluation Package - End-to-End**

Prices per package (for evaluation use only)

ELRA members: 500 Euro
Non-members: 750 Euro

Special discount for a combined purchase of several TC-STAR Evaluation Packages:

- Between 4 and 8 TC-STAR Evaluation Packages: -10%
- Between 9 and 16 TC-STAR Evaluation Packages: -20%
- Beyond 16 TC-STAR Evaluation Packages: -40%