Data Collection for the CHIL CLEAR 2007 Evaluation Campaign

N. Moreau[1], D. Mostefa[1], R. Stiefelhagen[2], S. Burger[3], K. Choukri[1]
[1]ELDA, [2]UKA-ISL, [3]CMU
Evaluations and Language resources distribution agency (ELDA)
55-57, rue Brillat Savarin 75013 Paris, France
E-mail: moreau@elda.org, mostef@elda.org, stiefel@ira.uka.de, sburger@cs.cmu.edu, choukri@elda.org

Abstract

This paper describes in detail the data that was collected and annotated during the third and final year of the CHIL project. This data was used for the CLEAR evaluation campaign in spring 2007. The paper also introduces the CHIL Evaluation Package 2007 that resulted from this campaign including a complete description of the performed evaluation tasks. This evaluation package will be made available to the community through the ELDA General Catalogue.

1. Introduction

The project CHIL [1] (as "Computers in the Human Interaction Loop") was an Integrated Project (IP 506909) funded by the European Commission under its 6th Framework Program. The project started in January 2004 and finished in August 2007.

The goal of CHIL was to develop computer assistants that attend to human activities, interactions, and intentions. The research consortium included 15 leading research laboratories from 9 countries representing today's state of the art in multimodal and perceptual user interface technologies in the European Union and the US.

For the CHIL research effort to be successful, it needed to be accompanied by rigorous evaluations of the developed technologies. This allowed performance benchmarking and a better understanding of possible limitations and challenging conditions. In 2005 it was decided to completely open up the project-internal evaluations. Hence, the open international evaluation workshop CLEAR[2] (as "CLassification of Events, Activities, and Relationships") was created. Parts of the CHIL technologies were evaluated in CLEAR (Stiefelhagen, 2006). Two CLEAR evaluation campaigns were conducted, one in spring 2006 and one in spring 2007.

A key enabler of the CLEAR evaluations was the availability of appropriate corpora, annotated with the necessary information. Thus a major outcome of the project has been the collection of a rich set of audiovisual material for each campaign. To serve development and evaluation of the CHIL technologies, multi-sensory audiovisual lectures and seminars were recorded inside smart rooms (CHIL rooms) at five different CHIL partner sites.

This paper describes in detail the data that was collected and annotated during the third and final year of the CHIL project. This data was used for the CLEAR evaluation campaign in spring 2007. The paper also introduces the CHIL Evaluation Package 2007 that resulted from this campaign including a description of the performed evaluation tasks.

2. Data Collection

During the initial years of the project, the CHIL consortium had recorded "non-interactive seminars" (lecture room scenario). In addition to non-interactive seminars, a few "interactive seminars" (conference room scenario) were collected for the CLEAR 2006 evaluations (Mostefa, 2006). This new recording scenario was able to accommodate new evaluations such as speaker activity detection and source localization (Stiefelhagen, 2006). Finally, for the CLEAR 2007 evaluation, the consortium decided to collect a brand new set of data, consisting exclusively of interactive seminars.

2.1. Interactive Seminars (Meetings)

The basic differences between lectures (non interactive seminars) and meetings (interactive seminars) are the number and setting of participants, their interactivity, and the addition of far-field microphone arrays and extensive usage of video in the lecture data collection.

The CLEAR 2007 data set was collected during the second half of 2006. Five of the CHIL partners recorded five high quality interactive seminars, each lasting at least thirty minutes. The number of people attending the seminars was set to be between three and six. The final data set consists of 25 multi-channel audiovisual recordings.

The collecting sites were located at different CHIL partner labs:

- AIT: Research and Education Society in Information Technologies at Athens Information Technology, Athens, Greece;
- IBM: IBM T.J. Watson Research Center, Yorktown Heights, USA;
- ITC-irst: Centro per la ricerca scientifica e technologica at the Instituto Trentino di Cultura, Trento, Italy;

[1] CHIL (Computers in the Human Interaction Loop): http://chil.server.de.

[2] CLEAR (Classification of Events, Activities, and Relationships Evaluation): http://www.clear-evaluation.org.

- UKA: Interactive Systems Labs of the Universität Karlsruhe, Germany;
- UPC: Universitat Politècnica de Catalunya, Barcelona, Spain.

In comparison to the previous years, the data collection process was improved by defining a common data quality standard for the entire CHIL consortium. The quality standard will be described in section 2.5.

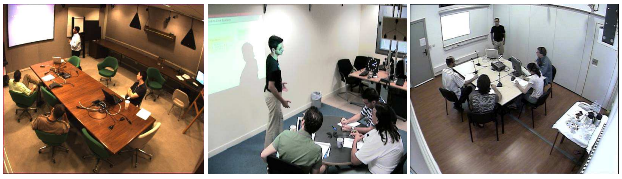
Each interactive seminar consists in a presentation given in a meeting room. These presentations are held by one or more speakers. The topics are related to technical matters of the CHIL project (mostly Natural Language Processing). The audience is small, between three and six people, and the attendees mostly sit around a table, all wearing close-talking microphones. There exists significant interaction between the presenters and the audience, with numerous questions and often a brief discussion among meeting participants. Typically, such scenarios include the following events:

- participants enter or leave the room,
- some attendees stand up and go to the whiteboard,
- discussions among the attendees,
- participants stand up for a short coffee break,
- during and after the presentation there are questions from the attendees with answers from the presenter.

In addition, a significant number of acoustic events are generated to allow more meaningful evaluation of the corresponding technology:

- sounds when opening and closing the door,
- interruptions of the meeting due to ringing mobile phones,
- attendees coughing and laughing,
- attendees pouring coffee in their cup and puting it on the table,
- attendees playing with their keys,
- keyboard typing, chair moving, etc.

Clearly, in such a scenario all participants are of interest to meeting analysis, therefore the CHIL corpus provides annotations for all (see Section 3). Examples camera views of interactive seminars are depicted in Figure 1.



IBM AIT UPC

Figure 1. Example camera views recorded at three CHIL smart rooms during meetings.

2.2. CHIL Smart Rooms

The five smart rooms are medium-size meeting or conference rooms that are equipped with a number of audio and video sensors, and have supporting computing infrastructure (Casas, 2004). The multitude of recording sites provides the desirable variability in the CHIL corpus, since the smart rooms obviously differ from each other in their size, layout, acoustics and visual environment (noise, lighting characteristics), as well as sensor properties (location, type) – see Figure 1. Nevertheless, it was crucial to produce a certain degree of homogeneity across sites to facilitate technology development and evaluations. Therefore, a minimum common hardware and software setup had been specified regarding the recording sensors and the data formats. All five sites complied with these minimum requirements, but frequently provided additional sensors. The minimum setup consists of:

- A set of common audio sensors, namely:
  - A 64-channel linear microphone array;
  - Three 4-channel T-shaped microphone clusters;
  - Three table-top microphones;
  - Close-talking microphones worn by the lecturer and each of the seminar participants.
- A set of common video sensors that includes:
  - Four fixed cameras located at the room corners;
  - One fixed, wide-angle panoramic camera mounted on the room's ceiling;
  - One active pan-tilt-zoom camera.

This sensor set is supported by a network of computers to capture the sensory data, mostly through dedicated data links. The data synchronization is realized in a variety of ways. A schematic diagram of such a room including its sensors is depicted in Figure 2.

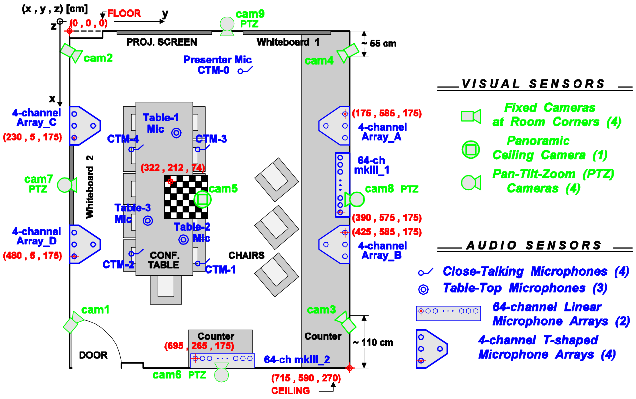


Figure 2. Schematic diagram of the IBM smart room, one of the five installations used for recording the data.

2.3. Audio Sensor Setup

Each smart room contains a minimum of 88 microphones that capture both closetalking and far-field acoustic data. In particular, for far-field audio recording, there exists at least one 64-channel linear microphone array, namely the Mark III array developed by NIST[3], placed on the smart room wall opposite to the speaker area. Such a sensor allows audio beamforming for speech recognition and speaker localization. The microphone array is accompanied by at least three additional microphone clusters located on the room walls, each consisting of four microphones organized in an inverted "T" formation of

[3] The NIST MarkIII Microphone Array: http://www.nist.gov/smartspace/cmaiii.html.

known geometry to allow far-field acoustic speaker localization. Additional far-field audio is collected by at least three table-top microphones. The latter are positioned on the meeting table, but their exact placement is not fixed. As a contrast to the far-field audio data, close-talking microphones are used to record the lecture presenter and, in the case of small meeting recordings, all the meeting participants. At least one of these microphones is wireless, to allow free movement of the presenter. Slight variations of this setup can be found among the five recording sites. For example, the IBM smart room contains two NIST Mark III arrays, whereas the ITC room has seven T-shaped arrays.

For audio data capture, all microphones not belonging to the NIST Mark III are connected to a number of RME Octamic eight-channel pre-amplifiers/digitizers.

The pre-amplifier outputs are sampled at 44.1 kHz and 24 bits per sample, and are recorded to a computer inWAV format via an RME Hammerfall HDSP9652 I/O card. The 64-channel NIST Mark III data are similarly sampled and recorded in SPHERE format, but are fed into a recording computer via an ethernet connection in the form of multiplexed IP packets.

2.4. Video Sensor Setup

The video data is captured by five fixed cameras. Four of them are mounted close to the corners of the room, by the ceiling, with significantly overlapping and wideangle fields-of-view. These are set in such a fashion, so that any person in the room is always visible by at least two cameras. The fifth camera is mounted on the ceiling, facing top-down, and uses a fish-eye lens to cover the entire room. The type of cameras installed varies among the sites, being either firewire or analog, providing images in resolutions ranging from 640 × 480 to 1024 × 768 pixels, and frame rates from 15 to 30 fps. All fixed cameras are calibrated with respect to a reference coordinate frame, with both extrinsic and intrinsic information provided in the corpus. In addition to the fixed cameras, at least one active pan-tilt-zoom (PTZ) camera is available in all five smart room setups. Its purpose is to provide close-up views of the presenter during lectures or meetings.

An example of smart room fixed camera views is depicted in Figure 3. For data capture, a number of dedicated computers are used, with all video streams saved as sequences of JPEG-compressed images. This allows easy nonlinear access to the frames, as well as exact absolute time stamping. It is also worth mentioning that most meeting recordings are accompanied by brief video sequences that contain empty room images captured immediately preceding the entry of all participants. These are provided to assist background modelling in video processing algorithms.

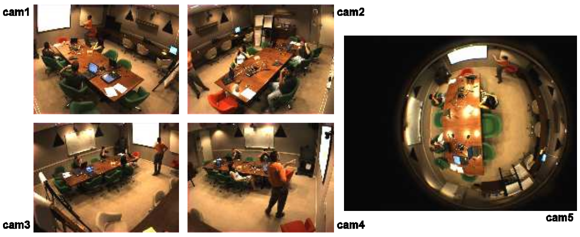


Figure 3. Sample synchronous images captured at the IBM smart room during an interactive seminar (meeting).

2.5. Quality Standard

In the beginning of the third year of the project, CHIL internally developed a new standard of quality for all sensors, in order to have each site producing the same quality of data, and to improve the data collection process set up in the previous years.

Video Quality Standard

Each site followed the recommendation of four angle cameras and a central ceiling mounted fish eye camera. The minimum frame rate was set to 15 frames per seconds (fps). The data streams were saved as sequences of JPEG images in a fixed name standard: seq xxxxx.jpg with xxxxx the number of the frame. A specific file called seq.index contained the table of correspondence between the frame and its associated time stamp. A file called seq.ini contained all the camera related information.

The maximum desynchronization between the five cameras for the entire length of a recording was set to 200ms. This was measured by introducing at the beginning and end of each recording a distinct and well observable audio-visual signal. The decision on how to realize this was left to the recording site but a movie studio-type clap was suggested. This was also a good way of testing the synchronization between the audio and video channels.

Microphone Array Quality Standard

Each site was equipped with at least one fully functional Mark III microphone array version 2. The version 2 was developed in collaboration with NIST. It generates 64 channels of audio, captured at 44 KHz and 24 bits of resolution. For each recording, the channel 4 was extracted. A specific file called timestamps.ini was created to store the time stamp of an eventual packet loss. The maximum desynchronization due to packet loss during one recording was fixed to 200ms. If more occurred, the recording had to be remade.

Hammerfall Quality Standard

Each site was equipped with at least 20 microphones for synchronous capture of audio. The former correspond to at least three T-shaped microphone arrays, each having four microphones, located on the walls. The remaining channels are from table-top microphones located on the conference table and close-talking ones. Just as for the MarkIII microphone array, a specific file called timestamps.ini was created to store the time stamp of an eventual packet loss. The maximum desynchronization due to packet loss during one recording was fixed to

50ms. If more occurred, the recording had to be remade.

Additional Information

To have every site providing the same information in a structured manner, a specific info directory was included in each recording. It contains a calibration directory with 10 pictures per camera and their calibration results and a background directory with pictures of the background before the meeting, when the room is empty.
A seminar datasheet was also required for each recording. It mainly contains information about the attendees: photo with identity tags, microphones corresponding to each attendee, etc. The presentation slides were also required. All this information was meant to make the transcription and annotation work easier and more reliable.

3. Data Annotations

As in the previous years, the CHIL 2007 corpus is accompanied by rich manual annotations of both its audio and visual modalities. Table 1 gives the amount of annotated data (lectures and meetings) produced for each of the CHIL evaluation campaigns.

Evaluation Campaign	Development Data	Evaluation Data
CHIL Internal	2h 20min	1h 40min
CLEAR06	2h 30min	3h 10min
CLEAR07	2h 45min	3h 25min

Table 1. Amount of annotated data for the CHIL /CLEAR evaluation campaigns.

3.1 Audio Channel Annotations

Data recording in the CHIL smart room results in multiple audio filescontaining signals recorded by close-talking microphones (near-field condition), table-top microphones, T-shaped clusters, and the Mark III microphone array (far-field condition), in parallel. The recorded speech as well as environmental acoustic events were carefully segmented and annotated by human transcribers at two locations, the European Language Resources Distribution Agency (ELDA) and the interACT Center at Carnegie Mellon University (CMU).

Orthographic transcriptions

Transcriptions were done by native English speakers with the Transcriber tool[4]. The manual transcription process started by transcribing the speaker contributions of all recorded near-field channels on orthographic word level, including the typical speaker-produced noises such as laughter and filled pauses. The start and end of the contributions were manually segmented. The transcription of the near-field condition was then compared to one of the far-field channels. Nonaudible events were removed and details recorded by only the far-field sensors were added.

These annotations were mainly used in the frame of the NIST Rich Transcription Meeting Recognition evaluation (RT 2007[5]) which was organized in cooperation with CLEAR 2007. The RT evaluation focused more on the evaluation of content-related technologies, such as speech and video text recognition. The evaluation data produced by CHIL for RT 2007 is described with more details in (Burger, 2007).

Annotation of Acoustic Events

Following the orthographic transcription of close-talking and far-field audio, a third step was performed for annotating environmental acoustic events. Such annotations were used in support of the “acoustic event detection and classification” task in the CLEAR evaluations.

Acoustic events describe all audible events in a recording. Accordingly, SPEECH is here also considered an acoustic event but is only broadly labeled as SPEECH, not transcribed in single words. Beside SPEECH, the set of labels for acoustic events consists of DOOR SLAM, STEP, CHAIR MOVING, CUP JINGLE, APPLAUSE, LAUGH, KEY JINGLE, COUGH, KEYBOARD TYPING, PHONE RINGING, MUSIC, KNOCK (door, table), PAPER WRAPPING, and UNKNOWN.

The annotation of acoustic events was carried out as an independent additional labeling process using the Annotation Graph Tool Kit (AGTK). Unlike Transcriber, AGTK enables the annotation of multiple overlapping events[6].

Acoustic events were labeled on two different types of data sets: acoustic events occurring in the CHIL lecture and meeting corpus and recordings of artificially produced events. The first set of data was labeled listening to the fourth channel of the Mark III microphone array. The artificially produced acoustic events were recorded in two data sets in the ITC and UPC smart rooms, and they contain isolated acoustic events collected in a quiet environment with no temporal overlap.

3.2 Video Channel Annotations

Facial Features and Head Location Information

Video annotations were manually generated using an ad-hoc tool provided by the University of Karlsruhe and modified by ELDA. The tool allows displaying one picture every second, in sequence, for all camera views. To generate labels, the annotator performs a number of clicks on the head region of the persons of interest, i.e., the lecturer only in the non-interactive seminar (lecture) scenario, but all participants in the interactive seminar (meeting) scenario. In particular, the annotator first clicks on the head centroid (e.g., the estimated center of the person’s head), followed by the left eye, the right eye, and the nose bridge (if visible). In addition, the annotator delimits the person’s face with a bounding box. The 2D coordinates of the marked points within the camera plane are saved to the corresponding label file. This allows the computation of the 3D head location of the persons of interest inside the room, based on camera calibration information. Figure 4 depicts an example of video labels, produced by this process. It shows the head centroid

[4] Transcriber Tool: http://trans.sourceforge.net

[5] The Rich Transcription 2007 Meeting Recognition Evaluation: http://www.nist.gov/speech/tests/rt/2007

[6] The AGTK Annotation Tool: http://agtk.sourceforge.net.

(white), the left eye (blue), the nose bridge (red), the right eye (green), and the face bounding box.

Head Pose Annotations

In addition to 2D face and 3D head location information, parts of the lecture recordings were also labeled with gross information about the lecturer's head pose. In particular, only eight head orientation classes were annotated, deemed to be a feasible task for human annotators, given the low-resolution captured views of the lecturer's head. The head orientation label corresponded to one of eight discrete orientation classes, ranging from a 0° to a 315° angle, with an increment of 45°. Overall, nineteen lecture videos were annotated with such information. These videos were used in the CLEAR head-pose technology evaluation.

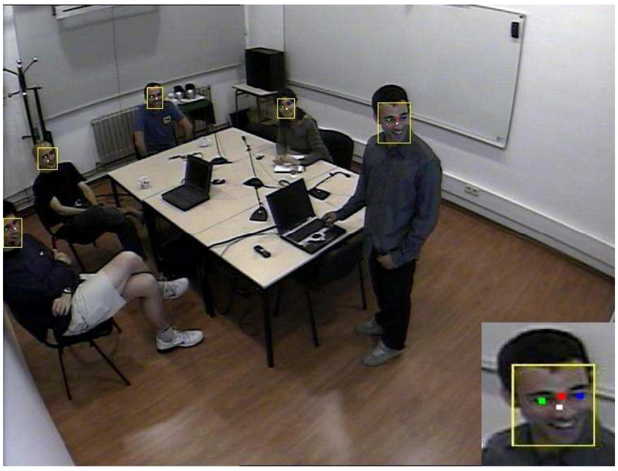


Figure 4. Example of video annotations for an interactive seminar in the UPC smart room.

3.3 Validation Procedures

The video annotations were validated internally. After being produced by human annotators, each annotation file was automatically scanned using a tool developed by ELDA. This tool detects most of the annotation errors that can occur: inversion of right and left eyes, missing labels, etc. During a second validation pass, a human operator checked and corrected manually the video labels. The error listings produced by the automatic scanning tool helped in this task. It was ensured that the person who checked a given seminar was different from the one who initially labeled it.

In the same way, each orthographic transcription was validated by a human transcriber, different from the one who produced it. A final pass was performed where all the data were reviewed by one person who used semi-automatic methods (spellchecker, lexicon, list of proper names, etc.) to check and correct the data. A further cross-validation check of video labels (at UKA) and audio transcriptions (between ELDA and CMU) was done. A few annotations were examined at random, to check if they were correct.

4. Evaluation Package

The CLEAR 2007 data sets have been made publicly available to the research community as part of the "CHIL 2007 Evaluation Package" (Moreau, 2007b) which is distributed by ELDA[7]. The technologies evaluated in CLEAR 2007 were:

- 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.

The complete evaluation package contains full documentation (definition and description of the evaluation methodologies, protocols and metrics) along with the data sets and software scoring tools, necessary to evaluate systems for each of the CLEAR 2007 technologies. Such a package therefore enables any developer to benchmark his systems and compare results to those obtained during the official evaluation. The CHIL 2007 Evaluation Package consists of the following:

- a document describing in detail the content of the package, as well as the corresponding evaluation (tasks, metrics, participants, results, etc.),
- the raw audio recordings of the seminars (Hammerfall, close talking microphones and microphone array channels),

[7] Evaluations and Language Resources Distribution Agency: http://www.elda.org

- the raw video recordings of the seminars (streams of the 4 corner cameras and ceiling camera),
- the video annotations and audio transcriptions of the seminars,
- useful information about each seminar (attendees, slides, calibration information, background pictures),
- additional databases specific to some evaluation tasks: Head Pose and Isolated Acoustic Events.

In addition, a range of specific data is provided for each evaluation task, allowing the package user to reproduce the evaluation in the same conditions:

- documentation about the evaluation procedure (metrics, submission format, etc.),
- the input data, as received by the participants during the evaluation,
- the participants' submissions,
- the reference labels,
- the scoring tools,
- the participants' results.

The CHIL 2007 Evaluation Package comes after the two previous CHIL packages released in 2005 and 2006.

5. Conclusion

A new evaluation data set has been produced for CLEAR 2007 during the 3rd year of the CHIL project. It consisted in recording interactive meetings through a large variety of audio and video sensors, in 5 different locations. This data set has been enriched by the manual annotations of both its audio and visual modalities.

The resulting CLEAR 2007 evaluation package (enclosing data sets, scoring tools and documentation) is publicly available to the community through the ELDA General Catalog[8]. Its goal is to enable external players to benchmark their system and compare their results with those obtained during the official evaluation campaign.

6. Acknowledgements

The data collection presented here was funded by the European Union under the integrated project CHIL (Grant number IST-506909).

7. References

Burger, S. (2007). *The CHIL RT07 Evaluation Data*. Rich Transcription 2007 Meeting Recognition Evaluation Workshop, May 2007.

Casas J., Stiefelhagen R. *et al.* (2004). *Multi-camera/multi-microphone system design for continuous room monitoring*, CHIL-WP4-D4.1-V2.1-2004-07-08-CO, CHIL Consortium Deliverable D4.1, July 2004.

Moreau N., Mostefa D., Stiefelhagen R. (2007a). *Perceptual Component Evaluation and Data Collection*, in: Alex Waibel, Rainer Stiefelhagen (Eds.), "CHIL: Computers in the Human Interaction Loop", Springer, 2007.

Moreau N. *et al.* (2007b). *Exploitation Material for the CHIL Evaluation Campaign 3*. CHIL Public Deliverable D7.14.

Mostefa D., Garcia M.-N., Choukri K. (2006). *Evaluation of Multimodal Components within CHIL*, in Proceedings of the 5th International Language Resources and Evaluations Conference (LREC 2006), Genoa, Italy.

Mostefa D., Moreau N. *et al.* (2007). The CHIL Audiovisual Corpus for Lecture and Meeting Analysis inside Smart Rooms, *Journal on Language Resources and Evaluation*, 41(3-4), pp. 389-407.

Stiefelhagen R., Bernardin K. *et al* (2006). *The CLEAR 2006 Evaluation,* Proceedings of the first International CLEAR evaluation workshop, CLEAR 2006, Springer Lecture Notes in Computer Science, No. 4122., pp 1-45.

[8] http://catalog.elda.org