The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

ELRA

January - December 2009

Vol.14 n.1 to 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.

Editors in Chief:
Victoria Arranz
Khalid Choukri

Editors:
Valérie Mapelli
Hélène Mazo

Layout:
Valérie Mapelli

Contributors:
Olivier Hamon
Nicolas Moreau
Djamel Mostefa

ISSN: 1026-8200

ELRA/ELDA
Secretary General: 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 special issue of the ELRA Newsletter is devoted to the evaluation campaigns and services that ELRA/ELDA have been promoting over the last decade. It is important that the community has a clear picture of the performance achieved by the tools, technologies and applications it develops. It is essential for the understanding of the field maturity but it is also critical to pave the way to new innovative approaches and paradigms.

The general mission of Human Language Technologies (HLT) evaluation is to enable the improvement of the quality of language engineering products. It is essential for validating research hypotheses, assessing progress and comparing research alternatives. It helps identifying mature technologies but also promising research directions. It also provides useful feedback to funding agencies and research managers.

Evaluation of Human Language Technologies has been applied in a number of technology areas and has proved efficient in providing adequate test-data collections, a better definition of profiles of users populations, classification schemes for HLT systems, a set of representative evaluation tasks, metrics for effectiveness, efficiency and satisfaction, but above all, a good landscape of the technology state of the art.

In our field, in addition to comparing different systems for a given application in a pragmatic way, evaluation aims at establishing the strong and the weak points of a system for further development.

Evaluation can be *user-oriented* when the users' satisfaction is used as an evaluation metric (very often related to the evaluation of advanced prototypes or ready-to-sell products although it is also used when automatic evaluations are not meaningful); this is very close to “**Usability evaluation**” that aims to measure the level of usability of a system. Typically, it enables users to achieve a specified goal in an efficient manner.

Research evaluation tends to validate new ideas or to assess improvements that new ideas bring in and this is what most of the evaluation campaigns are about. It is related to “**Performance evaluation**” that aims to assess the performance and relevance of a technology for solving a well-defined problem. When the processing chain of a system consists of several components associated at different stages, an additional distinction should be respected between *intrinsic* evaluation, designed to evaluate each component independently, and *extrinsic* evaluation meant to assess the overall performance of the system.

There are in general two main testing techniques for system measurement: *glass box* and *black box* which approximate the intrinsic and extrinsic evaluations. In the former, the test data is built by taking into account the individual component of the tested system. In the latter, however, the test data is chosen, for a given application, only according to the specified relations between input and output without considering the internal component.

Most of the technology evaluations are based on “Comparative evaluation” which is a paradigm in which a set of participants compare the results of their systems using the same data and control tasks with metrics that are agreed upon. Usually this evaluation is performed in a number of successive evaluation campaigns with open participation. For every campaign, the results are presented and compared in special workshops where the methods used by the participants are discussed and contrasted.

The experience with comparative evaluation in the USA and in Europe has shown that the approach has led to significant improvement of the performance of the evaluated technologies.

A number of key European projects have focussed on evaluation of Speech and Language Technologies and have led to a substantial literature in the area but also to good/best practices (although more practical evaluation campaigns have been conducted in the USA), such as EAGLES, ELSE (Bernsen, N. O., M. Blasband, N. Calzolari, J.P. Chanod, K. Choukri, L. Dybkjaer, R. Gaizauskas, S. Krauwer, I. de Lamberterie, J. Mariani, K. Netter, P. Paroubek, M. Rajman, A. Zampolli. 1999. *A Blueprint for a General Infrastructure for Natural Language Processing Systems Evaluation Using Semi-Automatic Quantitative Black Box Approach in a Multilingual Environment*, Deliverable D1.1 ELSE Project - Evaluation in Language and Speech Engineering LE4-8340, LIMSI, Paris.). For instance EAGLES (Expert Advisory Group on Language Engineering Standards) evaluation working group identified the 7 major steps for a successful evaluation. These steps, widely used today, are:

1. Why is the evaluation being done?
2. Elaborate a task model
3. Define top level quality characteristics
4. Produce detailed requirements for the system under evaluation, on the basis of 2 and 3
5. Devise the metrics to be applied to the system for the requirements produced under 4
6. Design the execution of the evaluation
7. Execute the evaluation

This newsletter elaborates on various Human Language Technologies with a focus on technologies in which ELRA/ELDA played a major role. These are selected as good examples of how an evaluation campaign can be set up: with strong involvement from the whole community, data and evaluation centres able to play an effective but neutral role, a scientific partner acknowledged within the community and bringing in scientific knowledge and know-how, as well as with a number of research participants from academia and industry. A number of details are given and perspectives for new evaluations are drawn.

Stelios Piperidis, President   Khalid Choukri, Secretary General

EVALUATION OF TECHNOLOGIES

Information Retrieval

Information retrieval (IR) refers to all technologies that aim at extracting and searching for information from large data collections. It is nowadays a key technology for knowledge management, guaranteeing access to large corpora of unstructured data. For instance, IR is the basic technology behind web search engines and an everyday technology for many web users.

Basically, IR systems allow a user to retrieve the documents which best match his information need(s) from a large document collection. The user's information need is expressed as a *query* (usually a few keywords taken from a natural language).

As depicted in Figure 1, an IR system deals with both the storage and representation of knowledge (*indexing*) as well as the retrieval of information relevant to a specific user's need (*search*). These steps in the functioning of an IR system can be described briefly as follows:

- *Indexing*: Document representations (*e.g.* sets of keywords) are extracted and stored.
- *Search*: The user's query is compared to the document representations.
- *Results*: The most similar documents (usually provided as a ranked list) are presented to the users who can evaluate the relevance with respect to their information needs and problems.

Figure 1 shows the most common IR framework, but other information access applications are addressed by the IR field, for instance, among others:

- *Question Answering (QA)*: information needs are expressed as natural language statements or questions. In contrast to classical IR where complete documents are considered relevant to the information need, a QA system returns specific pieces of information as an answer.
- *Information Filtering (IF)*: an information filtering system removes redundant or unwanted information from an information stream (e.g. identify spam and non-spam mails),
- *Topic Detection and Tracking (TDT)*: TDT refers to automatic techniques for finding topic-related materials in streams of data such as newswire and broadcast news.

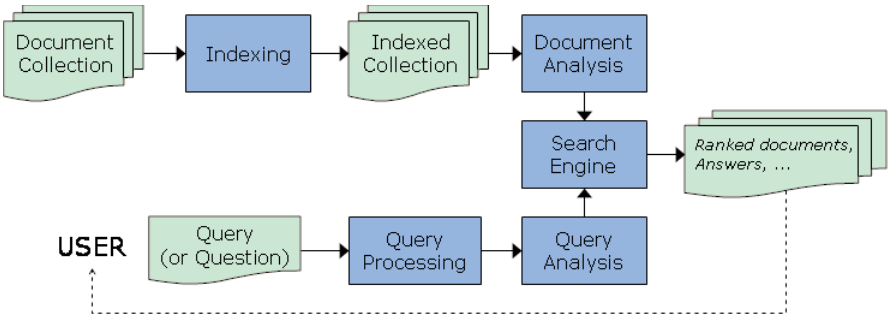

Figure 1: Basic structure of an IR system

One of the key issues nowadays in the IR domain is the development of Cross-Language Information Retrieval (CLIR) systems, which deals with collections containing multiple languages that are not necessarily the same as the language of the queries.

Another important issue is the development of multimodal IR technologies. In recent years, the growth of multimedia data (images with captions, video streams...) on the web and in professional and private databases has required the development of new information access strategies that combine features extracted from different modalities (text, audio, image, video, etc.).

Evaluation

The main IR evaluation campaigns carried out until now have adopted a comparative approach. Unlike objective evaluation (*How well does a method work?*), comparative evaluation focuses on the comparison of the results obtained with different systems (*Which method works best?*). To be compared, IR systems must be tested under similar conditions.

An IR comparative evaluation usually relies on a test collection consisting of:

- a set of indexed documents, which are to be searched (in order to identify those documents which satisfy the information needs),
- a set of test queries prepared from the test documents.

Once a test collection has been created, the general evaluation methodology is done in 3 main steps:

- *Evaluation run*: each IR system to evaluate searches the test collection using the pre-defined test queries. It yields a ranked list of document for each test query.
- *Relevance judgments*: human evaluators examine each retrieved document and decide if it is *relevant* or not, i.e. if it satisfies or not the information need expressed by the query.
- *Scoring*: performance measures are computed based on the human relevance judgments.

As long as they are tested on the same test collection (same set of documents and queries) the performance of different systems can be compared based on their final performance measures.

The performance of IR systems is usually measured by examining the first *N* retrieved documents and by computing:

- the percentage of documents that are relevant within the *N* top list (the *Precision* measure), and
- the percentage of all relevant documents that are included in the *N* top list (the *Recall* measure).

Precision and *Recall* values are computed for different values of *N*. The averaged *Precision* value is then used as a single indicator of retrieval efficiency.

Other specific IR tasks may require other performance measures may not be adequate. For example, the performance of QA systems is measured upon the percentage of correct answers obtained from the set of test questions.

The human relevance judgment step represents the most time- and resource-consuming part of an IR evaluation procedure:

- It requires the hiring of a team of objective experts who have to behave as if they were real users, and judge the relevance of each retrieved document with regard to the queries.
- A human evaluation framework (computer interface, evaluation guidelines, training sessions) must be carefully designed to ensure that all evaluators work under the same conditions.

Main Evaluation Campaigns

The Text REtrieval Conference (TREC[1]), was the first large-scale evaluation initiative for IR. It was started in 1992, co-sponsored by the US Defense Advanced Research Projects Agency (DARPA[2]) and the National Institute of Standards and Technology (NIST[3]). As a result of the success of these campaigns, some non-US agencies have begun to request that laboratories they fund test their systems within the DARPA evaluation framework.

As mentioned earlier, interoperability between languages is a key issue in the global information society. From 1997 to 1999, a track for the evaluation of cross-lingual IR (CLIR) systems was included in TREC. This track was coordinated jointly by NIST and a set of European volunteers, responsible for language-specific activities such as topic development and result assessment.

This CLIR issue is particularly important in the multilingual/multicultural European context. At the end of 1999, it was decided to move the cross-language evaluation activity for European languages to Europe. In 2000, an evaluation framework called Cross-Language Evaluation Forum (CLEF[4]) was set up within the DELOS Network of Excellence for Digital Libraries[5]. Since then, CLEF evaluation campaigns have been conducted on a yearly basis, addressing many different languages and tasks (IR, QA, Image retrieval, etc.)

In Japan, similarly to CLEF, the NTCIR project[6] adopted the TREC methodology and developed a multilingual evaluation framework for East Asian languages. The institution organizing the NTCIR evaluation is the National Institute for Informatics (NII) in Tokyo where the workshops have been held since the first campaign in 1997.

Many other local IR evaluation initiatives are or have been carried out to address specific languages. For instance, the EQueR (QA) and Amaryllis (IR) campaigns in France focused on the French language, and the Forum for Information Retrieval Evaluation (FIRE[7]) initiative was started in 2008 to provide test environments for the major languages spoken in India.

Other projects have focused on more particular aspects of IR. The Initiative for the Evaluation of XML Retrieval (INEX[8]) started in 2002. An evaluation campaign is annually organized to address the retrieval from documents structured in XML.

Some projects reflect the growing interest for multimedia information retrieval, such as TRECVid[9] (started in 2003, from a TREC evaluation track), the ImageCLEF and VideoCLEF evaluation tracks of the recent CLEF campaigns, or the Franco-German Quaero[10] initiative. An evaluation campaign has even been set up for music retrieval, the Music Information Retrieval Evaluation eXchange (MIREX[11]).

ELDA has participated, as main or co-organizer, in several evaluation campaigns in the IR field.

Launched at the end of 1995, the **Amaryllis** project aimed at evaluating IR software for French text corpora. We participated in the creation of document corpora, questions and answers for French, similarly to what was done for English within the TREC project.

Also within a French context, we conducted the evaluation for the **EQueR** project (Evaluation campaign for Question-Answering systems) which was part of the Technolangue programme, funded by the French Ministry of Research and New Technologies. This project was designed to provide an evaluation framework for Question-Answering systems working in French.

A solid and multilingual expertise was acquired in the evaluation of cross-lingual and multimodal IR technologies through its involvement in all **CLEF** (Cross-Language Evaluation Forum) campaigns, since the beginning of the CLEF project in 2000. Within CLEF, ELDA has been in charge of a number of best-practice surveys in the domain of CLIR evaluation resources, and of several data creation and evaluation tasks along the years, dealing with various IR and QA technologies in a multilingual context.

The quality and availability of the evaluation resources created during those different campaigns has also been our responsibility, comprising tasks such as validation, packaging and distribution of the data as test suites through the ELRA catalogue[12].

Through this long and rich experience, we have developed an expertise in the following activities:

- Identification of evaluation requirements for a specific IR technology. The goal is to make evaluations more real-world-oriented and increase its value for a specific application area. It is of critical importance to avoid building artificial test collections on which evaluated systems would perform well, while they would not necessarily work well in a real scenario.
- Creation of appropriate test collections. This implies both collecting and formatting a large set of test documents which are specific to the domain of the evaluation scenario, as well as preparing an adequate set of test queries.
- Recruitment and training of a team of experts to perform the required human judgment tests on IR or QA system outputs.
- Development of appropriate tools, such as scoring scripts and interfaces for human evaluations.

[1] http://trec.nist.gov/

[2] DARPA has organised other evaluation initiatives related to information access and understanding. Important examples include the MUC (Message Understanding Conference) series, and the activities of TIDES (Translingual Information Detection, Extraction and Summarization).

[3] http://www.nist.gov

[4] http://www.clef-campaign.org/

[5] http://www.delos.info/

[6] http://research.nii.ac.jp/ntcir/

[7] http://www.isical.ac.in/~clia/

[8] http://inex.is.informatik.uni-duisburg.de/

[9] http://www-nlpir.nist.gov/projects/trecvid/

[10] http://www.quaero.org/

[11] http://www.music-ir.org/mirex/2009

[12] http://catalog.elra.info/

Speech Recognition

For humans, speech is the most natural and immediate form of communication. Therefore, at the heart of any speech-enabled application lies a speech recognition module.

Speech is a complicated signal produced as a result of several transformations occurring at different stages: this may be at the linguistic, articulatory, acoustic levels, etc.

Speech recognition technologies refer to technologies that aim at extracting relevant information from speech output. We can divide speech recognition into the following categories:

- *Automatic speech recognition (ASR)*, also know as "speech-to-text", is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone.
- *Speaker recognition* consists in recognizing who is speaking.
- *Spoken language identification* is the process of determining which natural language can be found in some given speech data.
- *Spoken language understanding* is a process that allows not only to recognize the words that are spoken but also to interpret or understand them.

ASR, speaker recognition and language identification can be described as a pattern recognition problem. Systems are made of three main modules.

Firstly, a signal processing module converts the signal to acoustic features. This is done to reduce dimension and redundancy, remove irrelevant data and convert the raw speech into a parametric representation. Predominant analysis techniques are filter bank analysis and linear predictive analysis applied on 5 to 20 millisecond windows.

Secondly, pattern matching is used to compare the pattern to be recognized (converted signal) to reference patterns or models. Each model or reference template represents a speech phenomenon (e.g. a phoneme, a word, a speaker, a language). Statistical modeling techniques are widely used to train these models using techniques such as Hidden Markov Models, Gaussian Mixture Models, Artificial Neural Networks, etc.

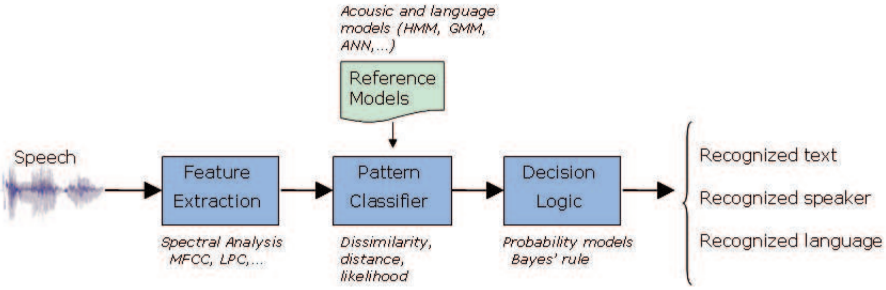
Acousic and language models (HMM, GMM, ANN,...)
Reference Models
Speech
Feature Extraction
Pattern Classifier
Decision Logic
Spectral Analysis MFCC, LPC,...
Dissimilarity, distance, likelihood
Probability models Bayes' rule
Recognized text
Recognized speaker
Recognized language

Figure 1: Basic structure for speech recognition technologies

Finally, a decision module is applied to identify the closest template by finding the model with the highest probability. The decision is usually made using *a posteriori* probability techniques and the Bayes' rule.

Figure 1 illustrates these three steps within speech recognition technologies, with details on the techniques applied for each module.

Evaluation

The quality of any speech application is critically dependent on the recognition performance of its speech recognition module. Moreover, quantitative measures of the performance of a speech recognition system are essential for making fundamental advances in the state-of-the-art. This fact is embodied in the statement, "You improve what you measure." In order to assess system performance, it is essential to have metrics or objective functions which are well-suited to the problem under investigation.

ASR Evaluation

Evaluation of ASR systems is mainly performed by computing the Word Error Rate (WER), or Character Error Rate (CER) for non Latin languages like Chinese or Japanese. The WER is also used for assessing other technologies like Machine Translation, for instance.

The WER is derived from the Levenshtein distance (or edit distance) and measures the distance between the hypothesis transcription produced by the ASR module and the reference transcription.

The WER is computed after the alignment between the hypothesis and the reference transcriptions has been done by dynamic programming (the optimal alignment being the one which minimises the Levenshtein distance). Once the alignment between the hypothesis and the reference is done, the WER counts the number of recognition errors.

Three kinds of errors are taken into account when computing the word error rate, i.e. substitution, deletion and insertion errors:

- Substitution: a reference word is replaced by another word in the best alignment between the reference and the system hypothesis.

- Deletion: a reference word is not present in the system hypothesis in the best alignment.

- Insertion: some extra words are present in the system hypothesis in the best alignment between the reference and the hypothesis.

The WER is defined as the sum of substitutions, deletions and insertions divided by the number of words in the reference transcription.

Although word is the basic unit for assessing ASR systems, the same computation can be made using different granularities (phonemes, syllables, etc.)

The WER can be higher than 100%, if the number of errors is more important than the number of words.

Prior to scoring both hypothesis and reference have to be normalised. The normalisation consists in converting the transcription into a more standardised form. This step is language dependent and applies a number of rules for transforming each token into its normalised form. For instance, numbers are spelled out, punctuation marks are removed, contractions are expanded, multiple orthographies are converted into a unique form, etc.

Although the WER is the main metric for assessing ASR systems, its major drawback is that all word errors are equally penalized, regardless of the importance and meaning of the word, e.g. an empty word has the same importance as a named entity.

Performance of ASR systems is also evaluated in terms of speed by measuring the processing time and computing the real time factor on a specific hardware configuration. This is an important factor for some applications that may require a real-time processing speed or for some devices that are limited in terms of memory or processor speed.

Speaker Recognition Evaluation

We distinguish **speaker identification** from **speaker verification**.

For **speaker identification**, systems have to give an identity to each test segment. The set of reference speakers is close and systems have to make a decision on whether a speaker is a specific person or is among a group of persons. Speaker identification can be text dependent or text independent.

In text-dependent identification, the phrase is known by the system and can be fixed or prompted (either visually or orally).

Assessment of speaker identification systems is performed by computing the "misclassification" rate.

For **speaker verification**, the task consists in deciding whether or not a specified speaker is speaking during a given speech segment. As in speaker identification, the speaker verification task can be performed in two ways, i.e. either text-dependently or text-independently.

When assessing speaker verification systems, two kinds of errors can be observed: (1) "false acceptance" when an impostor is accepted as the speaker he claimed to be and (2) "false rejection" when a correct speaker is rejected.

For each test hypothesis, both a detection decision and a score are required. The decision can be made by fixing a threshold on the test set *a posteriori*.

Various cost functions can also be used since one might think that false acceptance is more critical than false rejection. A particular functional point is the *equal error rate* computed under the constraint that false acceptance is equal to false rejection.

System capabilities are usually described with ROC and DET curves instead of using a single performance number. ROC stands for Receiver Operating Characteristic or, alternatively, Relative Operating Characteristic, and plots the performance of the system using false alarm rate on the horizontal axis and correct detection rate on the vertical axis.

DET or Detection Error Tradeoff rate curves are a linearized version of ROC curves. They plot system performance using FAR and FRR errors on both axes.

For both speaker identification and speaker verification, different evaluations can be carried out depending on the nature of the data, the length of each training/test segment and the match/mismatch condition. When the training and test data are recorded in the same acoustic conditions, we refer to "match condition evaluation". However, when the training and test data are recorded in different acoustic conditions, we refer to "mismatch condition evaluation".

Spoken Language Recognition Evaluation

Like speaker identification and verification, spoken language recognition is also a classification problem and the evaluation techniques are very similar to the ones described in the previous section. Language recognition can be evaluated either as an identification problem or as a detection task. In the former case, the system has to associate a language to a test segment, while in the detection scenario, given a test segment and a language hypothesis, the system has to give a Boolean decision. In the identification scenario, performance is measured according to the misclassification rate. In the detection scenario, ROC/DET curves are used and cost functions that combine false alarm and false rejection are computed.

Spoken Language Understanding Evaluation

Spoken language understanding systems can be broken down into three modules: a speech recogniser, a natural language analyser and a dialogue manager.

Performance assessment for spoken language understanding systems is substantially more complex and problematic than for ASR systems. There exist several approaches to evaluate speech understanding and dialogue systems. On the one hand, performance can be assessed through subjective evaluations. After using the system, human beings are asked to fill in a debriefing questionnaire whose answers are used in the data analysis and performance evaluation. This evaluation is an end-to-end evaluation and is very expensive. On the other hand, spoken language understanding systems can be assessed with the evaluation of each module. The speech recogniser is evaluated with WER as described earlier in this document. The natural language analyser is evaluated with the paradigm of the semantic Concept Error Rate (CER). Given an ontology, a speech utterance can be annotated with semantic segments. Then, the reference semantic annotation done by human beings and the hypothesis semantic annotation performed by the system are compared. The CER is then computed as the WER but using semantic concepts instead of words for word error rate. This requires an alignment between the reference and hypothesis, which is then followed by the counting of the number of substitutions, deletions and insertions. One drawback of this metric is that it is sensitive to the order of the segments. For instance, if the reference and system annotation of a given sentence contain the same semantic concepts but these are not in the same order, then the metric will count errors regardless of the fact that all semantic information is correct and present in the system hypothesis.

To overcome this order problem, Precision and Recall measures can also be applied for comparing the reference and the system annotations. Precision is the ratio of correct segments in the system hypothesis divided by the number of segments in system hypothesis while Recall is the number of correct segments divided by the number of segments in the reference.

The evaluation of the natural language analyser can be done either out-of-context or in-context. For out-of-context evaluation, each speaker turn is annotated individually and without taking into consideration the dialogue history. With regard to in-context evaluation, each speaker turn has to be annotated with semantic segments taking into account the dialogue history.

Main Evaluation Campaigns

Evaluation of Speech in the U.S.

Main evaluation campaigns were organized in the U.S. by the National Institute of Standards and Technology (NIST[1]) and sponsored by the U.S. Defense Advanced Research Projects Agency (DARPA[2]).

DARPA started a speech evaluation campaign in the early 80s and the first evaluations were run in 1987.

Since then, they are organizing almost annual competitive evaluations of speech technologies, including broadcast news transcription (1996-1999), conversational telephone recognition (1997-2001), rich transcription (2003-present), language recognition (1996-present), and speaker recognition (2003-present)[3].

For spoken dialogue systems, DARPA organized in 1989 the Air Travel Information System (ATIS) evaluation campaign for which performance of five spoken dialogue systems was assessed.

In 1999, the COMMUNICATOR project was launched. It was a multi-layer multi-site project on advanced spoken dialogue systems research.

Evaluation of Speech in Europe

In France, the Aupelf-Uref[4] (International Association of French Speaking Universities) launched the ARC program in 1994, which was based on the evaluation paradigm for both spoken and written language for 7 different control tasks including Voice Dictation and Vocal Dialogue evaluations.

In Europe, various influential projects have tried to build the foundations of en evaluation methodology for spoken dialogue systems, starting with the francophone project AUF-Arc B2 in 1994.

The ESPRIT SUNDIAL project ran for five years, concluding in August 1993. The objective of the project was to design and build telephone-access spoken language interfaces to computer databases. After introducing the aims and objectives of the project, the problems of specifying an interactive system were outlined and the Wizard-of-Oz simulation method described. The architecture of the resulting system was introduced, and system transaction success results of up to 96.6 % were reported. In the final section, some implications for machine translation - particularly interpretive telephony - were identified.

The French DEFI (“CHALLENGE”) evaluation of Spoken Language Understanding was organized by 4 French universities and aimed at defining a methodology with a high diagnostic power, as opposed to standard ATIS-like frameworks.

The DISC project (1997-1998), funded by the European Commission, aimed at developing a systematic and general scheme for in-depth characterization of practice in the development and evaluation of spoken dialogue systems.

Several evaluation campaigns in the field of speech recognition, speaker recognition or spoken language understanding have taken place in the framework of EU and national programmes/projects, with the participation of ELDA. The most important ones are described briefly below.

ESTER/ESTER-II

The aim of the ESTER[5] evaluation campaign series (2003-2009) was to evaluate automatic broadcast news transcription systems for French language. These campaigns implemented several tasks divided into three main categories: orthographic transcription, event detection and tracking (*e.g.* speech vs. music, speaker tracking) and information extraction (*e.g.* named entity detection). A database of 100 hours of speech broadcast news was produced, which included detailed orthographic transcriptions, speaker segmentation and named entities. The transcription task was evaluated with WER, the segmentation tasks were evaluated with the misclassification rate and the named entities recognition task was evaluated with the slot error rate.

MEDIA

The MEDIA[6] project (2003-2006) aimed at comparing and diagnosing the context-sensitive understanding capability of spoken language understanding systems and it was coordinated by ELDA. We produced a speech dialogue database collected through the Wizard-of-Oz technique in the hotel reservation domain. 1,250 dialogues were collected, transcribed and annotated with out-of-context and in-context semantic segments. The corpus was then used for evaluation adopting the Concept Error Rate as a main metric with different levels of constraints. Precision and Recall measures were also computed for the comparison.

CHIL

The CHIL[7] project (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 on January 1st, 2004 and ended in 2007. CHIL attempted to develop computer assistants that attend to human activities, interactions, and intentions, instead of reacting only to explicit user requests 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.

We were responsible for the coordination of the data collection and technology evaluation tasks of the project. Four evaluation campaigns were organized between 2004 and 2007 and for speech the following technologies were evaluated:

- Close-Talking Automatic Speech Recognition

[1] http://www.nist.gov/speech

[2] http://www.darpa.mil

[3] http://www.itl.nist.gov/iad/mig/tests/rt/

[4] http://www.limsi.fr/Recherche/FRANCIL/frcl.html

[5] ESTER is the French acronym for “Evaluation de Systèmes de Transcription enrichie d'Emissions Radiophoniques” (Evaluation of Radio Broadcast Rich Transcription Systems). http://www.afcp-parole.org/ester/

[6] MEDIA is the French acronym for “Méthodes d'Evaluation des systemes de DIAlogue” (Evaluation methods of dialogue systems). http://elda.org/article115.html

[7] http://chil.server.de/servlet/is/101/

- Far-Field Automatic Speech Recognition
- Acoustic Person Tracking
- Acoustic Speaker Recognition
- Speech Activity Detection
- Acoustic Event Detection

TC-STAR

The Speech-to-Speech Translation TC-STAR project[8] (2004-2007) targeted a selection of unconstrained conversational speech domains -speeches and broadcast news- and three languages: European English, European Spanish, and Mandarin Chinese. The long-term research goal of the project was effective speech-to-speech translation of unrestricted conversational speech on large domains of discourse.

To assess the advances in all technologies, annual competitive evaluations, including three ASR evaluation campaigns in English, Spanish and Mandarin Chinese, were organized by ELDA in 2005, 2006 and 2007, respectively. These evaluation campaigns were open to external participants.

Within the ASR evaluation campaign, the Character Error Rate was used for Chinese whereas the Word Error Rate was used for English and Spanish. The evaluation was done in a case sensitive mode and we also evaluated the correctness of the punctuation marks in the system output.

[8] http://www.tcstar.org/

Machine Translation

Machine Translation (MT) aims at converting a text written in a source language into a text written in a target language, using an automatic process. The beginning of Computational Linguistics in the early 50s was an attempt to move towards the automatic translation of documents, with the aim of sharing knowledge without language barriers. A considerable ground has been covered during these past 60 years although we are still far from the utopia of a fully automatic translation comparable in quality to that of a human translator.

Machine Translation is currently used in specific domains with a limited vocabulary (such as weather reports translated by the METEO system[1]) or as a help to provide a first-draft translation to be improved by means of editorial work. Moreover, with the Internet boom, more and more people need fast but not necessarily high-quality translation.

MT systems translate a given text by decoding its meaning in the source language and then re-encoding it into the target language. The overall complexity of the operation stands in the cognitive aspect of the meaning and the way the machine interprets it.

Four general approaches are used in machine translation:

- *Rule-based translation*, for which the translation rules are written manually.

- *Example-based translation*, for which existing translations are used as models for the new translation.

- *Statistical-based translation*, for which the system is trained on parallel corpora and re-encoding is done by using probabilities.

- *Hybrid machine translation*, which aims at combining and profiting from the strong points of the 3 aforementioned approaches.

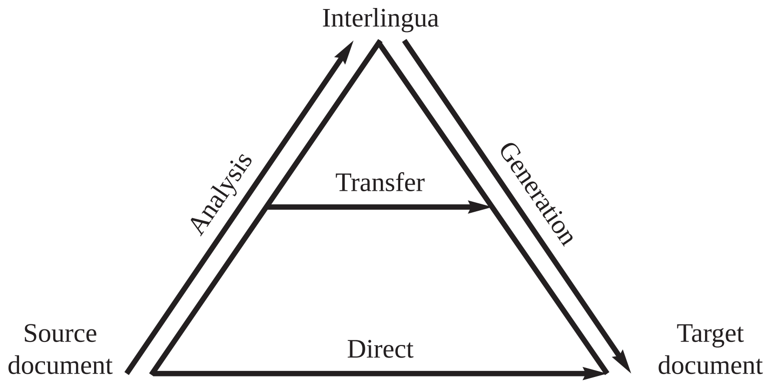

Figure 1: Different levels of linguistic analysis within an MT system

With regard to the linguistic complexity of the rule-based approach and its intermediary representation, three main methodologies can be distinguished. These are shown in Figure 1.

Generally, a machine translation system analyses a source document and then generates a target document. There are three possible approaches:

- *Direct translation*, which makes use of bilingual dictionaries.

- *Ttransfer translation*, which uses a source-language specific intermediate representation of the meaning that needs to be converted into target language by means of source-into-target transfer rules and generation rules.

- *Interlingual translation*, which uses an intermediate language-independent interpretation of the source text called *interlingua* that may be converted into the target language by means of generation rules.

MT Evaluation

A number of evaluation measures exist nowadays for evaluating MT systems but, generally speaking, two main methodologies can be distinguished: *human* and *automatic*. With human evaluations, the evaluator conducts judgements with the help of human judges (human subjects assessing the resulting translations). Each judge receives a certain number of sentences, randomly distributed among several judges. Judgements are then given according to

[1] Nirenburg, S. (1993). Progress in machine translation. IOS Press, Amsterdam/Oxford/Washington, DC./

evaluation criteria. The main criteria currently followed are *fluency* (“Is this text written in good target language?”) and *adequacy* (“Is the meaning of the candidate translation the same as that of the reference translation?”). The former aims at measuring the flow and naturalness of the target language, while the latter aims at measuring whether the translation keeps the meaning of the source text (by comparing the MT translation to a reference produced by a professional). Both criteria are usually evaluated on a 5-point scale and the average of all the judgements is computed once all judgements are done. Other kinds of judgements can also be processed, such as *preference*, that compares MT systems on the basis of one single sentence, or *informativeness*, that uses questionnaires to test the comprehension of the translation.

However, the use of human judgements is costly, due to the amount of work and the time needed for the judges to perform their judgements. To compensate for this cost, automatic evaluation has been introduced. Although its main drawback is the bad correlation with human judgement, automatic evaluation allows to have an idea on the MT system quality. The interpretation of the results according to the methodology and the metrics employed is thus extremely important.

Automatic measures usually compare a translation to one or more human reference translations, which are provided by professional agencies. Two trends are currently employed to compare two translations automatically. Most metrics use *n-gram* models (a consecutive sequence of *n* words, *n*=1…5) to compare sentences: the more n-grams the sentence has in common with the reference sentences, the higher the score. The original metric was BLEU (BiLingual Evaluation Understudy)[2], which adopts this single criteria of comparison. BLEU was then derived into several other metrics, such as:

- NIST (metric developed by the US National Institute of Standards and Technology): it introduces a penalty for much shorter candidates and the “information gain” that weights the rarity of an n-gram,
- WNM (Weighted N-gram Model): it refines the n-gram based comparison by weighting the words according to their importance,
- METEOR (Metric for Evaluation of Translation with Explicit Ordering): it uses stems and synonym matches, etc.

Other kinds of measures use error rates, based on WER (Word Error Rate) from the automatic speech recognition domain. This has also been derived into several metrics, such as:

- PER (Position independent Error Rate), which does not take word position into account,
- CER (Character Error Rate), which is an adaptation for Chinese, and
- HTER (Human Translation Error Rate), which measures the minimum number of manual edits required to change translation into one of the human references.

Main Evaluation Programs

Historically, the evaluation of MT systems has been directly linked to their development. In 1966, the ALPAC report[3] ordered by the US government stated the limitations of Machine Translation and the incapacity for computers to perform a “Fully Automatic High Quality Translation”. However, the report encouraged, in practice, the development of new approaches and, above all, constrained system developers to evaluate their results and to prove the good usage of funding. In 1979, the Van Slype report[4] studied a number of evaluation metrics in an MT context. But with the lack of evaluation campaigns, MT evaluation remained not standardized and scattered.

The first MT evaluation campaign, DARPA-MT, started in 1992. This 3-year evaluation was organized by the US Defense Advanced Research Projects Agency (DARPA[5]) and resulted in a new evaluation methodology in 1994. The 1994 campaign evaluated a dozen systems on the French-to-English direction.

In the meantime, the main goal of the JEIDA project[6] was to determine which MT system was more appropriate for the end-user. At the European level, the German project VerbMobil planned to develop an MT system that could translate in real time. A comparative evaluation of systems was then carried out.

After a six-year gap in the US, the National Institute of Standards and Technology (NIST[7]) took care of the organisation of a new series of annual evaluation campaigns in collaboration with the DARPA in 2001, and then by itself from 2006 onwards. In 2006 a new annual evaluation campaign also took place around the GALE project[8]. All in all, over a dozen evaluation campaigns have been organized since 2001. They have focused on a few translation directions like Chinese-to-English or Arabic-to-English, or more occasionally, on less-resourced languages such as Hindi-, Urdu- or Cebuano-to-English. Several evaluation conditions have been used, such as: evaluation with or without system development/training, evaluation on text or speech input, working with different topics or methodologies and metrics.

From 2004 to 2007, the European project TC-STAR[9] was set up with the aim of

[2] Papineni K, Roukos S, Ward T, and Wei-Jing Zhu 2001. BLEU: a Method for Automatic Evaluation of Machine Translation. IBM Research Division, Thomas J. Watson Research Center.

[3] ALPAC (1966). Languages and Machines: Computers in Translation and Linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council. (Publication 1416.).

[4] Van Slype, G. (1979). Critical Study of Methods for Evaluating the Quality of Machine Translation. Final report BR 19142, Brussels : Bureau Marcel van Dijk.

[5] DARPA has organised other evaluation initiatives related to information access and understanding. Important examples include the MUC (Message Understanding Conference) series, and the activities of TIDES (Translingual Information Detection, Extraction and Summarization).

[6] Nomura, H. (1992). Jeida Methodology and Criteria on Machine Translation Evaluation. Technical Report, Japan Electronic Industry Development Association (JEIDA), Tokyo.

[7] http://www.nist.gov

[8] http://www.itl.nist.gov/iad/mig/tests/gale/

[9] http://www.tcstar.org

developing a Speech-to-Speech Machine Translation system. Three evaluation campaigns took place during the project, on the three technologies involved: Automatic Speech Recognition, Machine Translation and Speech Synthesis. Three directions were used (Spanish-to-English, English-to-Spanish, Chinese-to-English) on data from the Spanish Parliament (for the former), from the European Parliament (for the former two) and on broadcast news (for the latter).

Several other evaluation campaigns have been carried out at a lower level, such as IWLST[10] (from 2005 to 2009) which evaluated MT systems on spontaneous conversation or read speech; WMT[11], initiated in 2005 and focusing on statistical MT, or the French CESTA project (from 2003 to 2007) which evaluated MT systems on general and specific domains for two language directions: English-to-French and Arabic-to-French.

Finally, the NIST introduced in 2008 the *NIST MetricsMATR*[12] for the Machine Translation Challenge, which aims at observing the performance of the evaluation metrics themselves and promoting the development of new metrics, and which organises a workshop jointly with WMT in 2010[13].

Two main projects have been carried out for the evaluation of MT systems at ELDA: CESTA and TC-STAR. As main organizers within the CESTA project, we run two evaluation campaigns for machine translation on the English-to-French and Arabic-to-French directions. Two kinds of input were used for the MT systems, one on a general domain and one on a specific domain (Health).

Being responsible for the whole evaluation process has implied the following:

- A Web interface has been developed for human judgements (reused for the TC-STAR evaluation).
- An evaluation protocol has been defined for the evaluation on French target data.
- Moreover, a *meta-evaluation* (evaluation of the evaluation metrics) has been conducted.

As co-organizers within the TC-STAR European project, we were responsible for the MT evaluation campaigns that were held between 2005 and 2007. Three language directions were considered within these campaigns (Spanish-to-English, English-to-Spanish and Chinese-to-English) as well as three different kinds of input (Final Text Editions provided by the European Parliament, manual transcriptions and automatic transcriptions).

All these evaluation campaigns helped in developing ELDA's evaluation expertise in the MT field. This covers all aspects, going from: the data preparation steps to the achievement of the results and their analysis. This implies that ELDA was in charge of:

- Identifying all potential resources to be collected, as well as
- Building the evaluation corpus to be prepared and formatted,
- Managing the whole translation process: once created, this evaluation corpus was translated by several translation agencies following the established specifications for the translation references,
- Hiring human judges so as to carry out human judgements,
- Collecting/creating tools and metrics to perform the evaluation,
- Computing evaluation scores and analysing results.

[10] http://www.is.cs.cmu.edu/iwslt2005/
[11] http://www.statmt.org/wpt05/
[12] http://www.itl.nist.gov/iad/mig/tests/metricsmatr/
[13] http://www.statmt.org/wmt10/

Speech Synthesis

Speech synthesis -often referred to as Text-To-Speech (TTS) processing- consists in converting written input to spoken output by automatically generating synthetic speech.

Although many processes are involved in a speech synthesis system, it mainly consists of the following three modules:

- Text Processing.
- Prosody Generation.
- Acoustic Synthesis.

Figure 1 depicts the basic structure of such a system and the coming sections detail the main modules involved.

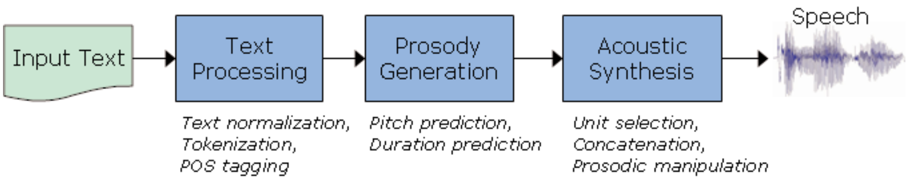

Figure 1: Basic structure of a TTS system

Text Processing

The first step in a TTS system is text processing. The input text is analyzed and transformed into a linguistic representation which contains all necessary information needed in the subsequent TTS steps. Some typical text processing tasks are the following:

- Identification of special words or symbols (numbers, acronyms, abbreviations, etc.) within the input text and their normalization (usually expansion in full text form).

- Assignment of a *part-of-speech* category (a *POS tag*) to each word from the input text. This POS determines its grammatical function.

- Setting up of a phonetic lexicon and a set of rules that are used to produce the appropriate phonetic transcription of the input text (*Grapheme-to-phoneme* conversion).

Prosody Generation

Prosody is a set of speech features that allows a same phonetic sound to be uttered in very different ways. These features include intonation (tone, pitch contour), speech rate, segment duration, phrase break, stress level and voice quality. Prosody plays a fundamental role to elicit the meaning, attitude and intention, and to produce natural speech.

The objective of the prosodic TTS module is to generate prosodic features that will make the intonation of the final synthesized speech as close as possible to a natural human voice intonation. For most TTS applications it is fundamental to produce expressive speech.

Acoustic Synthesis

The acoustic processing module physically generates the final speech signal (the *synthesized* voice) by implementing the appropriate sequence of phonetic units and the desired prosodic features.

The acoustic module exploits the textual (linguistic and phonetic) and prosodic information collected during the previous afore-mentioned processing stages.

TTS Evaluation

The quality of TTS systems can be assessed in two different ways:

- A first approach is to evaluate separately the components of the different TTS modules (glass box evaluation).
- Another (complementary) approach consists in measuring the global overall quality of a TTS system (black box evaluation).

TTS evaluation campaigns generally combine both approaches to investigate all objective and subjective aspects of speech synthesis technologies.

The complexity of TTS evaluation stands in the fact that it consists of separate evaluation tasks, each of them requiring a specific protocol and test collection. The four main TTS evaluation tasks are described in the following sections.

Text Processing Evaluation

Text processing evaluations require a large test collection of texts (typically 100k words). Each text processing module requires a particular evaluation task. Typically, the following basic processing modules are evaluated:

- *End-of-Sentence Detection*: segmentation of the text into autonomous sentences.

- *Normalization of Non-Standard-Words*: disambiguation of certain expressions such as abbreviations, dates, etc. (e.g. "*Mr*" changed to "*Mister*" or "29/02/08" to "*twenty ninth of February two thousand eight*").

- *POS Tagging*: labeling of each word with the appropriate grammatical tag.

- *Grapheme-to-Phoneme Conversion*: conversion of each word into the appropriate sequence of phonemes.

These evaluations are conducted separately by comparing the output of each module with reference data which have been manually produced by language experts. Comparisons are done automatically and result in error rates (percentage of incorrectly tagged words, segmented sentences, etc.) which enable to measure the performance of the different modules.

Prosody Generation Evaluation

This evaluation aims at assessing the ability of prosody generation systems to produce expressiveness features that will make the synthesized voice close to a natural human voice. The test collection is done as follows:

- A few test sentences are chosen.

- Sentences are manually normalized and phonetised by an expert, simulating the output of a perfect text processing module.

- All sentences are recorded by a speaker (baseline natural voice).

The evaluation procedure then consists in the following steps:

- Manually phonetised sentences are processed by the different prosodic modules to evaluate.

- Resulting prosodic descriptions of test sentences are collected.

- The baseline audio recordings (spoken by a natural speaker) are re-synthesized using the different prosodic descriptions. This results in different pronunciations of the same set of sentences that only differ in prosody (the baseline voice is the same).

- Subjective judgment tests are then performed by human listeners.

Each subject has to listen to different synthesized spoken sentences and to rate the naturalness of the voice (i.e. scoring it on a finite scale), paying attention to prosody only. The prosody generation modules under scrutiny are compared on the basis of these scores.

Acoustic Synthesis Evaluation

The test material is based on the test collection created for the prosody module evaluation (i.e. a few manually phonetised sentences and their recordings performed in a natural way by a human speaker). The next step is to add to these data a common syllabic and prosodic description (i.e. all information provided by a prosody description module). Prosodic descriptors indicate the duration, the fundamental frequency and the energy (or intensity) for each phoneme in the reference transcription. The evaluation procedure consists of the following steps:

- Test sentences and their prosodic descriptions are processed by the different acoustic modules to evaluate.

- Resulting synthesized speech excerpts are collected.

- Subjective judgment tests are then performed by human listeners.

Subjects are asked to rate the quality of the synthesized sentences they listen to, according to a series of pre-defined criteria (*naturalness*, *intelligibility*, *pleasantness*, etc.). The acoustic synthesis modules under scrutiny are compared on the basis of these scores.

Evaluation of Global TTS Quality

This evaluation takes the TTS processing chain as a whole, independently from its technological components (black-box evaluation). This time, the test collections simply consist of a few text sentences, with no additional information. The text sentences are processed by the TTS systems to evaluate. The resulting synthesized sentences are scored through the same kind of subjective tests as before (acoustic synthesis evaluation). The TTS systems under scrutiny are compared on the basis of these scores.

Main Evaluation Campaigns

A major contribution to the development of standardized TTS evaluation methodologies has been the TC-STAR [1] project (Technology and Corpora for Speech to Speech Translation) which ran from 2004 to 2007. Financed by the European Commission within the Sixth Framework Program, it was launched to support research advances in all core technologies for Speech-to-Speech Translation (SST): automatic speech recognition (ASR), spoken language translation (SLT) and speech synthesis (TTS). Core technologies were evaluated separately or in combination during the TC-STAR evaluation campaigns, focusing on English, Spanish and Chinese.

When the project ended in March 2007, the European Centre of Excellence in Speech Synthesis (ECESS[2]) was created to continue and extend the TTS activities of TC-STAR. ECESS is planning further TTS campaigns based on a web-based remote evaluation procedure.

There have been other national TTS evaluation initiatives in Europe, like the French EvaSy[3] project (part of the French Technolangue programme). EvaSy was run between 2003 and 2005 and performed speech synthesis evaluations for the French language.

In the United States, the major TTS project is CMU's Festvox[4], which has been organizing the yearly Blizzard Challenge[5] evaluation campaign since 2005.

Other important evaluation actions have addressed side aspects of speech synthesis, such as the HUMAINE[6] (Human-Machine Interaction Network on Emotion) FP6 European project, which took place from 2004 to 2007 and dealt with the expression of emotions in synthesized speech.

Regarding ELDA, it has participated, as main or co- organizer, in several evaluation campaigns in the TTS field.

The first project we were involved in was EvaSy an evaluation campaign for speech synthesis in French where we were responsible for the whole evaluation process. This was one of the very few evaluation experiences ever conducted on that topic for the French language.

Later, within the TC-STAR European project, which dealt with Speech-to-Speech Translation (SST) technologies, ELDA was in charge of the TTS part of the TC-STAR evaluation campaigns between 2005 and 2007. A large variety of evaluation tasks were addressed for TTS systems developed in three different languages: Chinese, English and Spanish. This resulted in the definition of new evaluation protocols for different aspects of the TTS research field.

During these campaigns, we:

- produced the test collections (data formatting and production of ground truth annotations) by settling partnerships with recognized expert linguists from different countries (UK, Spain, China). Produced data was based on the LC-STAR project where conventions were defined for phonetic transcriptions, POS tagging or syllabification.
- implemented and coordinated subjective evaluations for many different tasks thus acquiring a large know-how for setting up any kind of judgment tests for TTS technologies.

In the continuity of its TTS activities within TC-STAR, we were chosen to conduct the evaluations in the ECESS framework. ECESS aims at organizing TTS evaluation campaigns fully based on a web-based Remote Evaluation System (RES) platform (client-server architecture). Using the RES client, we have been able to perform different evaluation tasks via Web by sending test data directly to the modules or systems connected to RES and getting the results back.

These evaluation campaigns resulted in the creation of several TTS test data collections, including corpora, as well as associated annotations, tools, results, protocols etc. These packages are available to system developers for benchmarking purposes. The EvaSy and TC-STAR TTS packages are distributed via the ELRA catalogue[7]. ELDA is also responsible for the creation of future ECESS test suites. All these activities have given us a solid expertise in the creation, quality validation and distribution of TTS evaluation packages.

[1] http://www.tc-star.org/

[2] http://www.ecess.eu/

[3] http://www.technolangue.net/article.php3?id_article=202

[4] http://festvox.org/

[5] http://festvox.org/blizzard/

[6] http://emotion-research.net/

[7] http://catalog.elra.info/

Speech-to-Speech Translation

Speech-to-Speech Translation (SST) aims at translating a speech document recorded in a source language into another speech document recorded in a target language. So far, Speech-to-Speech Translation enables real-time translation for people who do not share a common language.

An SST system is composed of three chained modules that process the data trough an Automatic Speech Recognition (ASR) module, a Spoken Language Translation (SLT) module and a Text-to-Speech (TTS) processing module. Therefore, for a given speech document, an automatic transcription is produced by the ASR module in a source language, which is then sent to the SLT module to produce an automatic translation in a target language. Finally, this translation is synthesized in the target language by the TTS module in order to provide a speech document as output.

Figure 1 shows the module structure behind a Speech-to-Speech Translation system.

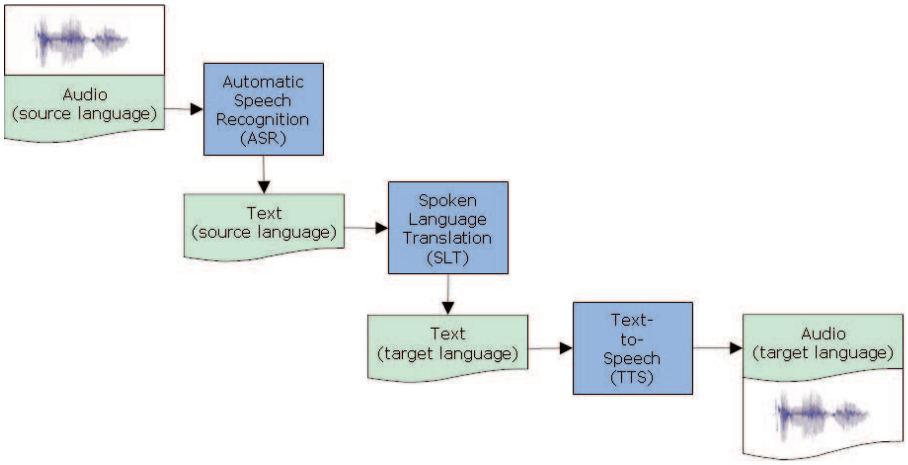

Figure 1: Basic structure of an SST system

The challenge in SST is thus to allow communication between people who do not speak the same language without the need to use human interpreters, who are not always available and who, from a practical point of view, remain a highly-demanding cognitive and expensive task. One of its objectives is then to reduce the costs of this task by using an automatic system.

SST systems may have some advantages, such as performing the translation of a long speech without needing to compress/summarise its meaning (like human interpreters generally do) or being able to reuse a system once it has been adapted to a specific domain. However, the main drawback of such a system still remains its output translation quality: current systems are still hard to understand and use, which proves that evaluation needs are all the more justified.

SST Evaluation

Evaluation in speech-to-speech translation jeopardizes many concepts and implies a lot of subjectivity. Three components are involved in the full translation process, which certainly increases the difficulty to estimate the output quality for an overall system. However, two criteria are mainly accepted within the community: measuring the information preserved during translation (transferred content) and determining how understandable the translation is (how well written the output is). Estimating the performance of an SST system implies focusing on End-to-End evaluation, where the evaluator considers the system as a whole and uses both input and output of the full system to proceed.

Most methodologies used within the projects depicted below differ in terms of data preparation, rating, or procedure.

Although the evaluation protocol mainly focuses on an end-to-end evaluation, we also consider the output from each module, namely the automatic transcription, the automatic translation and the speech synthesis. Their own evaluation, by following the same protocol, helps in detecting system issues and then allows for some kinds of improvement.

End-to-End evaluation for SST systems uses judges who generally rate speech output or answer questionnaires after listening to a speech document. Furthermore, they may also be asked to paraphrase what they understood. Nevertheless, the two basic concepts for this human evaluation are usually *adequacy* and *fluency*, also used in Machine Translation evaluation.

Adequacy evaluation is a comprehension test on potential users, which allows to measure the intelligibility (content) rate. The level of adequacy is computed as the percentage or number of questions that are assessed as correct. The objective is to know whether the speech-to-speech translation system is able to keep the meaning through the modules or not. To proceed, questionnaires are built by the evaluator, then filled in by the judges so as to observe the information loss or preservation. To that aim, the answers given by the judges are checked by an expert who may validate or reject them.

Fluency evaluation regards a judgement test with several questions related to fluency as well as the utility of the system. The objective here is to check whether the speech quality (form) of the output is sufficient. Fluency rates are based on a five-point scale or, less frequently, on a 3- or 7-point scale.

SST systems are also evaluated against professional interpreters. This is done by comparing their results with those obtained by the interpreters: for each document translated by the automatic system there is a corresponding document which has been interpreted and recorded by a human interpreter. Different judges are then asked to answer a questionnaire on the two outputs.

To carry out human evaluation, target-language native speakers are recruited. Depending on the task, they may be familiar or not with the speech-to-speech domain. They are also required to have no hearing impediments and should preferably not be bilingual in the source language in order to match the user task. Several speech documents are randomly given to the judges in such a way that each document is evaluated several times by different judges.

Main Evaluation Programs

Several end-to-end evaluations in speech-to-speech translation have been carried out in the last few years. One of the first ones was done within the JANUS project[1] in 1996, based on travel domain and using German and English as source languages, and English, German and Japanese as target languages.

This evaluation was followed in 1997 by the German Verbmobil project[2], which went on till 2000 and it worked on language directions such as German-to-English and German-to-Japanese.

Finally, one of the main projects to focus mainly on end-to-end evaluation was the TC-STAR project[3]. It provided two consecutive evaluations on English-to-Spanish translation. For further details on the activities within this project, please refer to the article on Machine Translation.

ELDA built its own expertise in the SST field through the organisation of two evaluation campaigns within the TC-STAR project. Within this project, ELDA was responsible for the SST evaluation campaigns between 2006 and 2007. Only the English-to-Spanish direction was evaluated within each campaign and only the TC-STAR system composed of an ASR combination of systems, an MT combination of systems (the best MT system for the first campaign) and a TTS system was evaluated.

Through these two campaigns:

- Evaluation tools were developed to perform human judgements and compute the results and scores.

- The guidelines and procedure were also defined for the resource production and the evaluation process.

Parsing

The goal of parsing is to represent the grammatical structure of a given text according to syntactic rules. This representation may take the form of a tree or some kind of hierarchical structure which represents the content of the analysed (or parsed) text. Generally speaking, from a given grammar, a parser analyses a text and provides its syntactic structure or syntactic tree. From this syntactic structure, the parser may establish the dependency relations that will help represent the meaning of the text during a semantic analysis. However, all this depends on the level and depth of analysis performed by

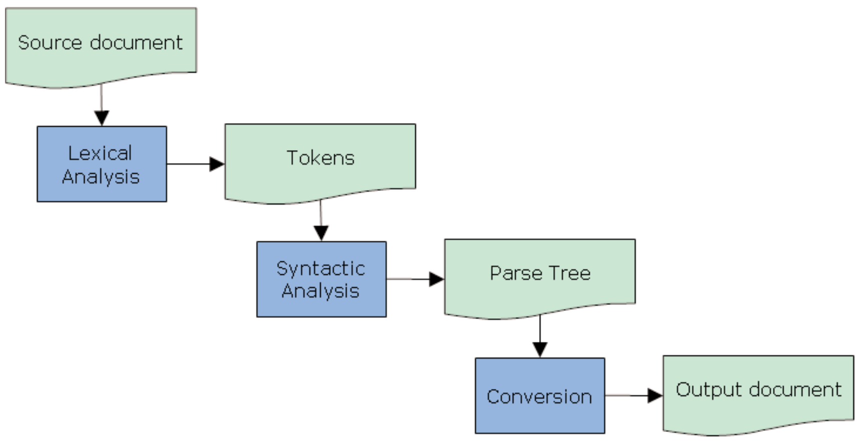

Figure 1: Basic structure of a parser

[1] Gates, D., Lavie, A., Levin, L., Waibel, A., Gavalda, M.,Mayfield, L. et Woszcyna, M. (1996). End-to-end Evaluation in Janus : A Speech-to-speech Translation System. In Proceedings of the 6th ECAI, Budapest.

[2] Nübel, R. (1997). End-to-end Evaluation in Verbmobil I. In Proceedings of the MT Summit VI, San Diego.

[3] http://www.tcstar.org

the parser. In this regard, we often refer to shallow parsing (which searches for the boundaries of major sentence elements such as noun phrases) or full parsing (aiming at establishing the relationships between the constituents and thus building up the full structure of the sentence.

The parsing approach followed depends on the type of grammar chosen. Different grammar formalisms exist (e.g., Lexical functional grammar, Head-driven phrase structure grammar, etc.) and the choice made will have an impact on the type of linguistic issues to be encountered and handled.

When looking at parsers in a larger-chain Natural Language Processing context, we could say that they are mainly used as primary technology to be integrated into other systems such as, historically, machine translation systems. This is also the case, for instance, of Part-of-Speech taggers, which in some NLP chains become the step preceding that of syntactic parsers.

For a general idea of the parsing process, please refer to Figure 1. As it can be seen, a traditional parser makes a lexical analysis of an input document so as to split it into meaningful symbols. A syntactic analysis is then carried out and a parse tree is built. Finally, the tree may be converted into a readable output document.

Parsing Evaluation

Parsers are evaluated by measuring the similarities between their automatic output and an annotated corpus which has been constructed manually by experts and which is used as reference.

As it happens with other technologies, the main distinction done in terms of evaluation methods regards whether these are useful for the development of a parsing system (which is generally referred to as *intrinsic evaluation*) or whether they are appropriate for comparing different systems (known as *comparative evaluation*).

Furthermore, parser evaluation methods can also be divided into corpus- and non-corpus-based methods, which is subject to the use or non-use of corpora, respectively, within the evaluation process.

The different methods can be listed and interrelated as follows[1]:

- Intrinsic evaluation:

- Listing linguistic constructions covered (no corpus used)
- Grammatical coverage (unannotated corpus)
- Average parse base (unannotated corpus)
- Structural consistency (annotated corpus)
- Best-first/Ranked consistency (annotated corpus)

- Comparative evaluation:

- Entropy/Perplexity (unannotated corpus)
- Part-Of-Speech assignment accuracy (annnotated corpus)
- Tree similarity (annotated corpus)
- Grammar evaluation interest group (GIEG) scheme (annotated corpus)
- Dependency structure-based scheme (annotated corpus)

Most of the parsing evaluation campaigns use comparative evaluation on annotated corpus following the Part-of-Speech assignment accuracy and the Dependency structure-based scheme.

The evaluation procedure is as follows: first of all, a large corpus is sent as input to the parser, which sends back as result the same corpus but automatically annotated with its syntactic information (e.g. with constituents and relations).

The metrics applied are automatic and use a manually-annotated reference corpus. The main issue here is to have a clear annotation formalism. When the output of a parser is available, it is compared to the reference so as to give different kinds of results, such as comparisons over the whole corpus or over different types of constituents or relations, etc. Results are generally given in terms of recall, precision or f-measure.

Main Evaluation Programs

Although there are not as many evaluation campaigns for parsing as for other domains, several have been carried out since the 90s.

In 1994, one of the first projects that included evaluation tasks for French parsers was GRACE[2] (*Grammars and Resources for Analyzers of Corpora and their Evaluation*), which ended in 1997. It allowed to establish a first standard for the annotation of syntactic constituents (e.g. nominal, adjectival, verbal) and dependency relations (e.g. subject/verb, direct object, adjective modifier) which was adopted by several parsers.

At the European level, the SPARKLE project[3] (*Shallow Parsing and Knowledge Extraction for Language Engineering*) started in 1997 and ended in 2000. Its main goal was to develop parsing technology for English, French, German and Italian. Evaluation was carried out using recall (number of correct constituents/relations found in relation to the number of correct constituents/relations to be found) and precision (number of correct constituents/relations found in relation to all constituents/relations found by the parser) measures.

In 1998, within the XTAG project[4] (wide-coverage grammar development project for English using a lexicalized Tree Adjoining Grammar (TAG) formalism), an evaluation task was carried out on English corpora using recall, precision and f-measure.

From 2003 to 2006, the EASY project[5], within the French national programme Technolangue, carried out an evaluation campaign on French by reusing the lessons

[1] Carroll, J., Briscoe, T., Sanfilippo, A. (1998). Parser Evaluation: a Survey and a New Proposal. In Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC), Granada, Spain.

[2] http://www.limsi.fr/RS99FF/CHM99FF/TLP99FF/tlp10

[3] http://www.ilc.cnr.it/sparkle/sparkle.html

[4] http://www.cs.sfu.ca/~anoop/papers/pdf/eval-final.pdf

[5] http://www.limsi.fr/Recherche/CORVAL/easy/

learnt from GRACE's experience. EASY demonstrates that there are many parsers available for French. This project has been followed by the PASSAGE project[6] which is still ongoing and extending the goals of the EASY project.

ELDA's expertise in the field of parsing has been developed around two main French projects that carried out a number of evaluation campaigns.

The EASY project was responsible for an evaluation campaign dedicated to parsing for French language. The project was launched within the French national programme Technolangue and ended in December 2006. One evaluation campaign was carried out involving twelve parsers. Beyond evaluation, the objectives of the project were 1) to define and validate an evaluation methodology to compare parsers for French and 2) to produce and validate manually a large annotated corpus by combining the output from the parsers automatically.

The PASSAGE project, which started in 2007 and is still running, follows the EASY project and aims at extending the initial goals to a large-scale production of syntactic annotations. Therefore, the main objective is to build semi-automatically a French Treebank of more than 100 million words by combining the output of several parsers and correcting it manually afterwards.

This long-term effort permits to get an expertise in parsing evaluation and also results in the collection of annotated corpora, the development of tools, and the support of an evaluation campaign. EASY and PASSAGE have also allowed the establishment of a large network of partners that are involved at several stages of the evaluation and parsing processes. We have certainly benefited from this additional expertise.

Last but not least, an evaluation server has been built as a Web service that allows the automatic evaluation of the systems via Internet. System developers can thus upload the results of their system being run on a defined corpus and obtain scores after a certain time delay.

[6] http://atoll.inria.fr/passage/

Multimodal Technologies

Multimodal technologies refer to all technologies combining features extracted from different modalities (text, audio, image, etc.). This covers a wide range and diversity of component technologies:

- Multimodal Information Retrieval,
- Audiovisual Speech Recognition,
- Audiovisual Person Identification,
- Audiovisual Event Detection,
- Audiovisual Object or Person Tracking,
- Head Pose Estimation,
- Gesture Recognition,
- Document Recognition (OCR),
- Biometric Identification (based on voice, fingerprints, iris, etc.), etc.

One of the most researched fields is multimodal information retrieval, which deals with IR techniques combining search in text, images, videos, audio documents etc. Two popular fields of research are Video Retrieval (based on key-frames, audio streams, textual annotations etc.) and Content-Based Image Retrieval (based on image features, OCR-ized text, textual annotations, etc.).

Evaluation

There is no generic evaluation approach for such a wide and heterogeneous range of technologies. In some cases, the evaluation paradigm is basically the same as for the equivalent mono-modal technology (e.g. traditional IR vs. multimodal IR). For very specific applications (e.g. 3D person tracking in a particular environment), *ad hoc* evaluation methodologies have to be pre-defined before the start of the evaluation campaign.

A good example of this is the multimodal evaluation framework set up for the CHIL project[1] (Computers in the Human Interaction Loop). Different test collections (production of ground truth annotations) and specific evaluation metrics were defined to address a large range of audio-visual technologies:

- Acoustic speaker identification & segmentation
- Acoustic emotion recognition
- Acoustic event detection
- Speech activity detection
- Face and Head tracking
- Visual Person tracking
- Visual Speaker Identification
- Head Pose Estimation
- Gesture Recognition
- Multimodal Person Identification
- Multimodal Person Tracking, etc.

A complete overview of these evaluation tasks can be found in the book that was published at the end of the project[2].

Main Evaluation Campaigns

Multimodal technologies encompass a large variety of technologies. Many past and present projects and evaluation campaigns address multimodal information processing.

In the recent years, the European Commission has funded large projects in the field of human machine interaction technologies. The CHIL project, in particular, organized two large evaluation campaigns called CLEAR (Classification of Events, Activities and Relationships) in 2006 and 2007[3], covering a large range of multimodal technologies (2D and 3D person tracking, head pose estimation, multimodal person identifi-

[1] http://chil.server.de

[2] Waibel, A. and Stiefelhagen, R. (Ed.): Computers in the Human Interaction Loop, Springer London, 2009.

[3] CLEAR 2006: http://isl.ira.uka.de/clear06/ and CLEAR 2007: http://isl.ira.uka.de/clear07/

cation, etc.). AMI[4] was another important EC-funded project (FP6), which like CHIL aimed at developing technologies that help people have more productive meetings. A follow-up of AMI, AMIDA, is currently carried on in the FP7.

Outside Europe, the most well-known evaluation project is VACE[5]. VACE is a US program including evaluations of object detection and video tracking technologies.

All these projects are dealing with both audio and visual information. Other projects cover a large range of technologies addressing the same modality, like TECHNO-VISION for visual issues. For instance, TECHNO-VISION was a French program that included several vision-related evaluation campaigns:

- ARGOS[6]: evaluation campaign for surveillance tools of video content,
- EPEIRES[7]: evaluation of symbol recognition methods,
- ETISEO[8]: video surveillance,
- EVALECHOCARD: medical imaging,
- IMAGEVAL[9]: image processing technology assessment
- IV2[10]: biometric iris and face identification,
- MESSIDOR[11]: methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology,
- RIMES[12]: evaluation campaign for handwritten document processing,
- ROBIN[13]: evaluation of object recognition algorithms,
- TOPVISION[14]: submarine imaging systems.

In the field of information retrieval, some large well-known projects reflect the growing interest for multimodal and multimedia information access, such as TRECVid[15] (started in 2003, from a TREC evaluation track), the ImageCLEF and VideoCLEF evaluation tracks of the recent CLEF campaigns, or the Quaero[16] initiative (a Germano-French collaborative research and development program, centered at developing multimedia and multilingual indexing and management tools). An evaluation campaign has even been set up for music retrieval, the Music Information Retrieval Evaluation eXchange (MIREX[17]).

ELDA has participated, as the main or co- organizer, in the international evaluation campaigns of the European project CHIL (cf. previous sections). During three years, the CHIL corpus has been the cornerstone in the evaluation of a multitude of audiovisual perception technologies for human activity analysis during lectures and meetings inside smart rooms. The following is a non-exhaustive list of some of the evaluated technologies covered:

- Person localization and tracking technologies,
- Person identification,
- Face recognition,
- Speaker identification,
- Gesture recognition,
- Conversational large-vocabulary continuous speech recognition,
- Acoustic scene analysis,
- Emotion identification,
- Topic identification,
- Head-pose estimation,
- Focus-of-attention analysis,
- Question answering, and summarization, etc.

In collaboration with other CHIL partners, ELDA designed a specific evaluation framework (submission format, metrics, scoring scripts, etc.) for each of the addressed technologies. In addition, we produced ground truth material for some tracks (in particular video annotations), and we were in charge of conducting the CHIL evaluation campaigns.

ELDA is also an active partner of the CLEF framework (Cross-Language Evaluation Forum) and has been involved in all CLEF evaluation campaigns since 2000. CLEF deals with cross-lingual information access in general, and with some cross-lingual and multimodal IR technologies in particular. It includes a cross-language image retrieval task, known as ImageCLEF, which deals with the issues involved in retrieval from an image collection when the user queries are expressed in a language different from that of the collection. Retrieval is based on low-level features derived from an image, or on the associated caption, or on a combination of both.

The evaluation resources created in CHIL and CLEF were validated and packaged by ELDA and they are distributed as test suites through the ELRA catalogue[1].

Through this long and rich experience, we have developed an expertise in the following activities:

- Creation of multimodal test collections (in particular audiovisual recordings).
- Development of specific tools for the annotation of images and /or video frames.
- Development of appropriate scoring scripts to compute the *adhoc* evaluation metrics (e.g. to compare the output of a face detection system with the ground-truth image).
- Packaging of multimodal resources.

[4] AMI (Augmented Multi-party Interaction): http://www.amiproject.org/

[5] VACE (Video Analysis and Content Extraction): http://www.perceptual-vision.com/vt4ns/vace_brochure.pdf

[6] http://www.irit.fr/PERSONNEL/SAMOVA/joly/argos/index.html

[7] http://epeires.loria.fr

[8] http://www-sop.inria.fr/orion/ETISEO/index.htm

[9] http://www.imageval.org/

[10] http://lsc.univ-evry.fr/techno/iv2/PageWeb-IV2.html

[11] http://messidor.crihan.fr/index-en.php

[12] http://rimes.it-sudparis.eu/

[13] http://robin.inrialpes.fr/

[14] http://topvision.gesma.fr/

[15] http://www-nlpir.nist.gov/projects/trecvid/

[16] http://www.quaero.org/

[17] http://www.music-ir.org/mirex/2009

Evaluation Services at ELRA/ELDA

Based on our know-how, a number of evaluation services are offered at different levels for a wide range of technologies. According to their needs, technology developers will select the type or level of service where we can help:

Production of evaluation specifications and guidelines

Input: ***Technology developer's evaluation need(s)***

For instance:

In order to search your web site or your own enterprise database:

- Do you want to benchmark an IR system developed internally or acquired outside?
- Do you want to compare different systems to find the technical solution that best fits your needs?

If you are an MT developer, a user, within a company or working as an individual, the range of evaluation topics is very broad, starting from the evaluation of a single product to the comparison of several MT systems. ELDA:

- can run an evaluation campaign.
- offers the benchmark of MT systems regarding the automatic translations produced.
- benefits from its neutral and independent position on the market to guarantee the objectivity of the results and their interpretation.

Capitalising on its experience in speech recognition evaluation, through its participation in different projects, ELDA can offer its expertise evaluating the following speech technologies:

- Automatic speech recognition: broadcast news transcription, close-talking microphone speech recognition, far-field speech recognition, telephone speech recognition, etc.
- Speaker recognition: text dependent or text independent speaker identification and verification.
- Speech segmentation: speaker diarization, music segmentation, speech segmentation.
- Language recognition: language identification and language verification.
- Spoken language understanding: subjective evaluation of human-machine dialogue systems, semantic concepts evaluation.
- Acoustic event detection.
- Acoustic source localization.

Process: ***Defining with the* technology developer *what the most appropriate evaluation scheme is according to his needs***

By closely examining the user's specific needs, we bring in our expertise for:

- Determining the most adequate evaluation strategy (e.g. building an evaluation framework from scratch or finding and using an existing evaluation package which suits the user's needs).
- Defining the appropriate resources to be collected or developed (i.e. nature of the documents, size of the data, ...).
- Budgeting the work to be done.
- Producing the evaluation specifications and guidelines to carry out the evaluation required.

For this, ELDA:

- may propose different evaluation scenarios, if several options fit the user's needs, but a final proposal will be submitted after discussion.
- will base its proposal on its own guidelines, which have been improved throughout its different campaigns and evaluation activities obtaining highly reliable documents, and which are revised and adapted to meet the technology developer's needs.

The user will be allowed to supervise or enquire about the progress of the evaluation at all times.

Output: ***Evaluation specifications and guidelines***

ELDA provides all necessary guidelines and specifications regarding the appropriate evaluation protocol, data collections and tools (or the appropriate off-the-shelf evaluation package, if this exists).

Production of evaluation data

Input: ***Evaluation specifications and guidelines***

Process: ***Data acquisition and/or production***

We are able to evaluate technologies either by using an already-existing evaluation package or by building a complete evaluation framework from scratch. Following the user's specifications, we:

- produce the necessary evaluation data (by finding existing data sources or creating them from scratch, setting up partnerships, clearing IPR issues, formatting data, etc.).
- develop the required evaluation tools (scoring tools, evaluation interfaces).

Output: ***Evaluation package***

The complete set of material necessary to perform the required evaluation (produced data or resources, human judgement interfaces, scoring tools,...) is produced and handed in.

IR

- A test document collection is produced.
- From the collected documents, test queries are created that simulate typical users' needs with regard to the target application (e.g. search in web data, in enterprise database, etc.).

SST

Our know-how stands in the managing of the full evaluation chain of an SST system. Through our activities in the field of SST evaluation and related components (Automatic Speech Recognition, Machine Translation, Text-to-Speech), we can deal with the end-to-end evaluation of a system as well as the separate evaluation of its components. These parameters need to be taken into account in order to plan for the specific steps of the evaluation.

Once the data collection needs have been specified (i.e. topic, amount of data, etc.), several time-consuming tasks are taken care of, such as:

- Recording of audio data.
- Manual transcriptions.
- Human translations through our network of professional translation agencies (for some projects, not only the original speech but also the audio interpreter's translation may be required, so as to compare system with human translation quality).
- Preparing questions and answers for human judgement, translation of questions and answers into the target language and building of questionnaires.
- Recruitment of a team of judges, composed of experts and/or end-users, and their training so as to perform human judgements, namely in answering comprehension questionnaires and some quality questions.

ASR

The language resources needed for the evaluation are produced. This includes the test set but also the development data and if needed the training data.

MT

The evaluation corpus can be produced, which implies the following:

- Collecting the data (the *source documents*).
- Producing the corresponding translations (the *reference translations*). As the evaluation metrics usually compare a translation to one or more human reference translations, we work with a network of independent professional translation agencies.
- Formatting and cleaning the data.

TTS

- Creation of test collections for testing TTS modules, ensuring the recording of baseline voices and the setting up of partnerships with expert linguists to produce ground truth annotations in any language.
- Setting up and/or development of the required evaluation tools, in particular of the man-machine interfaces for subjective listening tests.
- Development of the required scoring scripts.

Parsing

- Specifications are defined with the user, for parameters such as constituents and relations, topic, corpus, amount of data, etc. This is basic information to start the identification of potential existing data and/or the production of test data collections (i.e. annotated data and test corpus).
- For the production task, annotators need to be recruited who are experts in the field and specially trained for the task.

Multimodal Technologies

Most multimodal technologies require the design of a specific evaluation framework:

- Production of specific annotations, possibly including image or audio-visual annotations, to set up ground truth data,
- Definition of specific metric and scoring scripts to compare systems' output with ground truth data.

IR

The assessors are guided to perform the relevance judgements in a controlled evaluation environment, using the required interfaces.

ASR

- Developing the required software for running the evaluation. This can be:
  - simple evaluation scripts or
  - a more complicated wed-based architecture for remote evaluation over the Internet.
- Assessing the performance by using automatic tools (automatic evaluation).
- Running subjective evaluation campaigns where human beings are requested to do the assessment (manual evaluation).

Input: ***Ready-to-use evaluation package***

Process: ***Evaluation run***

We are able to evaluate technologies either by using an already-existing evaluation package or by building a complete evaluation framework from scratch (cf. the output Evaluation package mentioned above).

In both cases:

- The whole evaluation protocol is implemented.
- A team of competent assessors may be recruited and trained, if required by the evaluation.
- Both the processing of relevance judgements and the scoring of systems are performed with the ad-hoc tools.

MT

- Both human and automatic evaluations can be carried out.
- When evaluation is performed manually, we:
  - are responsible for the recruitment and training of experts or end-users to play the role of judges. This is done following the evaluation specifications such as the topic, the amount of data, the number of judgements required, etc. so as to perform correctly the human judgements.
  - can adapt a Web interface for human judgements that has been developed at ELDA for new evaluation tasks.
- When the evaluation is automatic, we:
  - install and/or develop the different evaluation metrics for the task.
  - execute the metrics and make sure the procedure is run correctly.

Procedure:

- The automatic translations are collected, then checked to be well incorporated into the evaluation tools (i.e. automatic metrics or evaluation interfaces for human judgements).
- Once all judgements and/or automatic results are performed, several scores are given according to the precision level required in the specifications.
- Then, the automatic and/or human results are checked and validated.
- Finally, the results are analysed and interpreted so as to produce a final report including all the parameters of the evaluation.

TTS

Process: Evaluation run

The evaluation procedure depends on the TTS module to assess:

- Automatic scoring for text processing modules (3a).
- Listening tests for the other TTS modules: prosody generation, acoustic synthesis and global TTS quality (3b).

(3a): Text processing modules (objective evaluations)

The test collection(s) are processed by the system(s) under evaluation.

System(s) outputs are formatted and scored against the annotations of reference (ground truth) using the appropriate tools.

(3b): Listening tests (subjective evaluations)

The test sentences are synthesized by the system(s) under evaluation. The resulting pieces of audio are rated by human listeners through subjective tests.

- The evaluation platform is implemented according to predefined criteria (e.g. minimum number of ratings to collect for each system, number of listeners to recruit, distribution of test sentences among listeners, etc.).
- A team of appropriate listeners is recruited and trained. Recruitment is done according to strict predefined criteria (e.g. age, native speakers only, no hearing impediments, no expertise in synthesized speech, etc.).
- Listeners are coordinated and guided to perform the listening tests in a controlled environment (similar headphones, quiet place, etc.) and by following the same procedure.

SST

- We have the expertise in developing evaluation tools specific to SST tasks in order to perform human judgements and compute the results and scores.
- Evaluation criteria are proposed, such as the number of judges, the amount of words or hours of audio, the number of judgements per sample, the definition of questionnaires, etc.

Parsing

- A parsing evaluation server has already been built as a Web service that allows the automatic evaluation of systems via Internet. System developers can then upload their results on a defined corpus and obtain evaluation scores.
- When data are processed by the parser(s), the results are sent through our evaluation server so as to be compared to a reference corpus which has been annotated by our experts. This allows to either rely on us to carry out the full evaluation service (from the use of parser(s) results up to their analysis), or to access our online parsing evaluation service in order to perform one's own evaluations, at will.
- Some post-processing may be also carried out, such as the production and validation of the output corpus by combining automatically the output from the parsers and correcting it manually.

Multimodal Technologies

In some cases, the evaluation feature will be the same as for the equivalent mono-modal technologies (e.g. Information Retrieval). Based on its expertise, ELDA is able to define and implement an evaluation framework for any kind of multimodal technology.

If the system is evaluated using an existing evaluation package, results are compared to those obtained during the past official campaigns (benchmarking of systems, compared to state-of-the-art systems tested during past campaigns).

Output: ***System(s) performance analysis and final report***

We compute system(s) scores, analyse the results and produce a final report.

An evaluation close-up meeting may be organised to present the results or a larger event, such as a workshop, to discuss the results openly, if so wished by the user.

Evaluation Packages available at ELRA/ELDA

AURORA Project Databases

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

AURORA-CD0002 AURORA Project Database 2.0

The Aurora project 2.0 is a revised version of the Noisy TI digits database to follow on the work of ETSI. This CD set is a replacement for the previous set (version 1.0 consisted of 2 CDs while version 2.0 now consists of 4 CDs) . This database is intended for the evaluation of algorithms for front-end feature extraction algorithms in background noise but may also be used more widely by speech researchers to evaluate and compare the performance of noise robust speech recognition algorithms.

AURORA-CD0003-01 AURORA Project database - Subset of SpeechDat-Car - Finnish database

This database is a subset of the SpeechDat-Car database in Finnish language which has been collected as part of the European Union funded SpeechDat-Car project. It contains isolated and connected Finnish digits spoken in different driving conditions inside a car.

AURORA-CD0003-02 AURORA Project database - Subset of SpeechDat-Car - Spanish database

This database is a subset of the SpeechDat-Car database in Spanish language which has been collected as part of the European Union funded SpeechDat-Car project. It contains isolated and connected Spanish digits spoken in different noise and driving conditions inside a car.

AURORA-CD0003-03 AURORA Project database - Subset of SpeechDat-Car - German database

This database is a subset of the SpeechDat-Car database in German language which has been collected as part of the European Union funded SpeechDat-Car project. It contains isolated and connected German digits spoken in different noise and driving conditions inside a car.

AURORA-CD0003-04 AURORA Project database - Subset of SpeechDat-Car - Danish database

This database is a subset of the SpeechDat-Car database in Danish language which has been collected as part of the European Union funded SpeechDat-Car project. It contains isolated and connected Danish digits spoken in different noise and driving conditions inside a car.

AURORA-CD0003-05 AURORA Project database - Subset of SpeechDat-Car - Italian database

This database is a subset of the Italian SpeechDat-Car database which has been collected as part of the European Union funded SpeechDat-Car project. It contains contains 2200 Italian connected digit utterances divided into training and testing utterances in different noise and driving conditions inside a car.

AURORA-CD0004-01 AURORA Project Database - Aurora 4a

The Aurora project has released a number of list files for performing the training and testing on the Wall Street Journal (WSJ0) data at two sampling rates -8 kHz and 16 kHz. The Aurora 4a database is based on the WSJ0 with artificial addition of noise over a range of signal to noise ratios. It contains both clean and multicondition training sets and 14 evaluation sets with different noise types and microphones.

AURORA-CD0004-02 AURORA Project Database - Aurora 4b

The Aurora project has released a number of list files for performing the training and testing on the Wall Street Journal (WSJ0) data at two sampling rates -8 kHz and 16 kHz. The Aurora 4b, has been released. It contains noisy versions of the Nov'92 WSJ0 development set.

TC-STAR Evaluation Packages

TC-STAR is a European integrated project focusing on Speech-to-Speech Translation (SST). To encourage significant breakthrough in all SST technologies, annual open competitive evaluations were organized. Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) Text-To-Speech (TTS) were evaluated independently and within an end-to-end system (E2E).

ELRA-E0002 TC-STAR 2005 Evaluation Package - ASR English

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

ELRA-E0003 TC-STAR 2005 Evaluation Package - ASR Spanish

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

ELRA-E0004 TC-STAR 2005 Evaluation Package - ASR Mandarin Chinese

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

ELRA-E0005 TC-STAR 2005 Evaluation Package - SLT English-to-Spanish

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

ELRA-E0006 TC-STAR 2005 Evaluation Package - SLT Spanish-to-English

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

ELRA-E0007 TC-STAR 2005 Evaluation Package - SLT Chinese-to-English

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

ELRA-E0011 TC-STAR 2006 Evaluation Package - ASR English

This package includes the material used for the TC-STAR 2006 Automatic Speech Recognition (ASR) second evaluation campaign for the English language. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

ELRA-E0012-01 TC-STAR 2006 Evaluation Package - ASR Spanish - CORTES

This package includes the material used for the TC-STAR 2006 Automatic Speech Recognition (ASR) second evaluation campaign for the Spanish language within the CORTES task. It includes 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.

ELRA-E0012-02 TC-STAR 2006 Evaluation Package - ASR Spanish - EPPS

This package includes the material used for the TC-STAR 2006 Automatic Speech Recognition (ASR) second evaluation campaign for the Spanish language within the EPPS task. It includes 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.

ELRA-E0013 TC-STAR 2006 Evaluation Package - ASR Mandarin Chinese

This package includes the material used for the TC-STAR 2006 Automatic Speech Recognition (ASR) second evaluation campaign for the Mandarin Chinese language. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

ELRA-E0014 TC-STAR 2006 Evaluation Package - SLT English-to-Spanish

This package includes the material used for the TC-STAR 2006 Spoken Language Translation (SLT) second evaluation campaign for English-to-Spanish translation. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

ELRA-E0015-01 TC-STAR 2006 Evaluation Package - SLT Spanish-to-English - CORTES

This package includes the material used for the TC-STAR 2006 Spoken Language Translation (SLT) second evaluation campaign for Spanish-to-English translation within the CORTES task. It includes 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.

ELRA-E0015-02 TC-STAR 2006 Evaluation Package - SLT Spanish-to-English - EPPS

This package includes the material used for the TC-STAR 2006 Spoken Language Translation (SLT) second evaluation campaign for Spanish-to-English translation within the EPPS task. It includes 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.

ELRA-E0016 TC-STAR 2006 Evaluation Package - SLT Chinese-to-English

This package includes the material used for the TC-STAR 2006 Spoken Language Translation (SLT) second evaluation campaign for Chinese-to-English translation. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

ELRA-E0025 TC-STAR 2007 Evaluation Package - ASR English

This package includes the material used for the TC-STAR 2007 Automatic Speech Recognition (ASR) third evaluation campaign for the English language. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

ELRA-E0026-01 TC-STAR 2007 Evaluation Package - ASR Spanish - CORTES

This package includes the material used for the TC-STAR 2007 Automatic Speech Recognition (ASR) third evaluation campaign for the Spanish language within the CORTES task. It includes 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.

ELRA-E0026-02 TC-STAR 2007 Evaluation Package - ASR Spanish - EPPS

This package includes the material used for the TC-STAR 2007 Automatic Speech Recognition (ASR) third evaluation campaign for the Spanish language within the EPPS task. It includes 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.

ELRA-E0027 TC-STAR 2007 Evaluation Package - ASR Mandarin Chinese

This package includes the material used for the TC-STAR 2007 Automatic Speech Recognition (ASR) third evaluation campaign for the Mandarin Chinese language. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

ELRA-E0028 TC-STAR 2007 Evaluation Package - SLT English-to-Spanish

This package includes the material used for the TC-STAR 2007 Spoken Language Translation (SLT) third evaluation campaign for English-to-Spanish translation. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

ELRA-E0029-01 TC-STAR 2007 Evaluation Package - SLT Spanish-to-English - CORTES

This package includes the material used for the TC-STAR 2007 Spoken Language Translation (SLT) third evaluation campaign for Spanish-to-English translation within the CORTES task. It includes 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.

ELRA-E0029-02 TC-STAR 2007 Evaluation Package - SLT Spanish-to-English - EPPS

This package includes the material used for the TC-STAR 2007 Spoken Language Translation (SLT) third evaluation campaign for Spanish-to-English translation within the EPPS task. It includes 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.

ELRA-E0030 TC-STAR 2007 Evaluation Package - SLT Chinese-to-English

This package includes the material used for the TC-STAR 2007 Spoken Language Translation (SLT) third evaluation campaign for Chinese-to-English translation. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

ELRA-E0031 TC-STAR 2006 Evaluation Package - End-to-End

This package includes the material used for the TC-STAR 2006 evaluation campaign within the end-to-end task. It includes 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.

ELRA-E0032 TC-STAR 2007 Evaluation Package - End-to-End

This package includes the material used for the TC-STAR 2007 evaluation campaign within the end-to-end task. It includes 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.

CHIL Evaluation Packages

The CHIL Evaluation Packages were produced within the CHIL Project (Computers in the Human Interaction Loop), in the framework of an Integrated Project (IP 506909) under the European Commission's Sixth Framework Programme. The objective of this project is to create environments in which computers serve humans who focus on interacting with other humans as opposed to having to attend to and being preoccupied with the machines themselves. “Instead of computers operating in an isolated manner, and Humans [thrust] in the loop [of computers] we will put Computers in the Human Interaction Loop (CHIL)”.

ELRA-E0009 CHIL 2004 Evaluation Package

The CHIL Seminars are scientific presentations given by students, faculty members or invited speakers in the field of multimodal interfaces and speech processing. The language is European English spoken by non native speakers. The recordings comprise the following: videos of the speaker and the audience from 4 fixed cameras, frontal close ups of the speaker, close talking and far-field microphone data of the speaker's voice and background sounds.
The database consists of:
1) Audio and Video Recordings of 10 seminars
2) Video annotations done displaying 1 over 10 pictures in sequence, for the 4 cameras.
3) Transcriptions using both TRS and STMUID formats.

ELRA-E0010 CHIL 2005 Evaluation Package

The CHIL Seminars are scientific presentations given by students, faculty members or invited speakers in the field of multimodal interfaces and speech processing. The language is European English spoken by non native speakers. The recordings comprise the following: videos of the speaker and the audience from 4 fixed cameras, frontal close ups of the speaker, close talking and far-field microphone data of the speaker's voice and background sounds.
The database consists of:
1) Contents of the CHIL 2004 Evaluation Package (see catalogue reference ELRA-E0009 for description).
2) Audio and Video Recordings: 5 seminars recorded in November 2004).
3) Stereo Video Recordings of 10 subjects that move in the camera's field of view while performing pointing gestures.
2) Video annotations.
3) Transcriptions.

ELRA-E0017 CHIL 2006 Evaluation Package

The CHIL Seminars are scientific presentations given by students, faculty members or invited speakers in the field of multimodal interfaces and speech processing. The language is European English spoken by non native speakers. The recordings comprise the following: videos of the speaker and the audience from 4 fixed cameras, frontal close ups of the speaker, close talking and far-field microphone data of the speaker's voice and background sounds.
The CHIL 2006 Evaluation Package consists of:
1) A set of audiovisual recordings of seminars, called non-interactive seminars and of highly-interactive small working groups' seminars, called interactive seminars. The recordings were done between 2004 and 2005 according to the “CHIL Room Setup” specification.
2) Video annotations.
3) Orthographic transcriptions.

ELRA-E0033 CHIL 2007 Evaluation Package

The CHIL Seminars are scientific presentations given by students, faculty members or invited speakers in the field of multimodal interfaces and speech processing. The language is European English spoken by non native speakers. The recordings comprise the following: videos of the speaker and the audience from 4 fixed cameras, frontal close ups of the speaker, close talking and far-field microphone data of the speaker's voice and background sounds.
The CHIL 2007 Evaluation Package consists of:
1) A set of audiovisual recordings of interactive seminars. The recordings were done between June and September 2006 according to the “CHIL Room Setup” specification.
2) Video annotations.
3) Orthographic transcriptions.

Technolangue Evaluation Packages

The Technolangue Evaluation Packages were produced within the French national research programme Technolangue funded by the French Ministry of Research and New Technologies (MRNT). Each package includes the material that was used and/or produced for each evaluation campaign: resources, protocols and metrics, scoring tools, results of the campaign, etc. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results against the state of the art.

ELRA-E0018 ARCADE II Evaluation Package

The ARCADE II project enabled to carry out a campaign for the evaluation in the field of multilingual alignment. The campaign is distributed over two actions:
1) Sentence alignment: it consists in evaluating the alignment of French language with Latin-script languages on one side, and with non Latin-script languages on the other side.
2) Translation of named entities: it consists in identifying in the parallel Arabic corpus the translation corresponding to the named entities phrases annotated in the French corpus.

ELRA-E0019 CESART Evaluation Package

The CESART project enabled to carry out a campaign for the evaluation of terminology extraction tools. This project is an extension of the evaluation campaign of terminology resource acquisition tools that was carried out for written corpora (ARC A3) within the AUPELF campaigns (Actions de recherche Concertées, 1996-1999). The campaign is distributed over two actions:
1) Term extraction for the building of a terminology reference which applications are the enrichment of the reference and the free indexing of documents.
2) Extraction of semantic relations (synonymy) from a list of “focal” terms.

ELRA-E0020 CESTA Evaluation Package

The CESTA project enabled to carry out a campaign for the evaluation of machine translation systems with English and Arabic texts translated into French. The campaign is distributed over two actions:
1) Evaluation on a restrictive vocabulary: an evaluation protocol was introduced and was dedicated to two translation directions: English into French and Arabic into French.
2) Evaluation on a specialised domain (evaluation after terminology enrichment): it consists in observing the impact of the systems adaptation to the specialised domain.

ELRA-E0021 ESTER Evaluation Package

The ESTER project enabled to carry out a campaign for the evaluation of Broadcast News enriched transcription systems using French data. This project is an extension of the only campaign that was ever carried out for French in this field within the AUPELF campaigns (Actions de recherche Concertées, 1996-1999). The campaign is distributed over three actions:
1) Orthographic transcription: it consists in producing an orthographic transcription of radio-broadcast news, which quality is measured by word error rates. There are two distinct tasks, one with and one without calculation time constraint.
2) Segmentation: the segmentation tasks consist of segmentation in sound events, speaker tracking and speaker segmentation. For the sound event segmentation, the task consists of tracking the parts which contain music (with or without speech) and the parts which contain speech (with or without music). The speaker tracking task consists in detecting the parts of the document that correspond to a given speaker. The speaker segmentation consists of segmenting the document in speakers and grouping the parts spoken by the same speaker.
3) Information extraction: it consists of an exploratory task on named entity tracking. The objective was to set up and test an evaluation protocol instead of measure performances. The systems must detect eight classes of entities (person, place, data, organisation, geo-political entity, amount, building and unknown) from the automatic transcription or the manual transcription.

ELRA-E0022 EQueR Evaluation Package

The EQueR project enabled to carry out a campaign for the evaluation of Question-Answering systems in French. The campaign is distributed over two actions:
1) Generic task: it consists in evaluating the performances of question-answering systems on a collection of heterogeneous texts.
2) Specialised task: it consists in evaluating the performances of question-answering systems on a collection of texts from the medical domain.

ELRA-E0023 EvaSy Evaluation Package

The EvaSy project enabled to carry out a campaign for the evaluation of speech synthesis systems using French text data. This project is an extension of the only campaign that was ever carried out for French in this field within the AUPELF campaigns (Actions de recherche Concertées, 1996-1999). The campaign is distributed over three actions:
1) Evaluation of grapheme-to-phoneme conversion: it consists in evaluating the capacity of speech synthesis systems to phonetize text data.
2) Evaluation of prosody: it consists in evaluating the capacity of speech synthesis systems to forecast text prosody (duration and fundamental frequency of phonemes) from the text itself.
3) Global evaluation of the quality of speech synthesis systems through ACR tests (Absolute Category Rating) and SUS tests (Semantically Unpredictable Sentences).

ELRA-E0024 MEDIA Evaluation Package

The MEDIA project enabled to carry out a campaign for the evaluation of man-machine dialogue systems for French. The campaign is distributed over two actions:
1) Evaluation taking into account the dialogue context: it consists in producing semantic annotation outside dialogue context for each of the 3,000 test prompts.
2) Evaluation not taking into account the dialogue context: it consists in evaluating the capacity of understanding systems a) from orthographic transcriptions only and b) from transcriptions and reference annotations outside dialogue context.

ELRA-E0034 EASy Evaluation Package

The EASy project enabled to carry out a campaign for the evaluation of syntactic parsers of French. The campaign is distributed over two actions:
1) Evaluation of constituent annotation: it consists in evaluating the ability of parsers with respect to the type of corpus (e.g. literature, conversation transcription, parliamentary speech, questions for information retrieval tools).
2) Evaluation of dependency relation annotation: it consists in evaluating the ability of parsers with respect to the relations between constituents or words.

Amaryllis Project

ELRA-W0029 Amaryllis Corpus

Launched at the end of 1995, the AMARYLLIS project aimed at evaluating information retrieval software for French text corpora in order to provide a methodology for the evaluation of other similar tools. AMARYLLIS was organised by the Institut de l'Information Scientifique et Technique (INIST) with the support of the *Agence francophone pour l'enseignement supérieur et la recherche (AUPELF-UREF)* and the French *Ministère de l'Education Nationale, de la Recherche et de la Technologie (MERT)*.
More specifically, the objective was to create document corpora, questions and answers, in the framework of the Action de Recherche Concertée (ARC A1, renamed as Amaryllis- Access to text information in French), in order to get similar works to the United States project TREC. All corpora are structured as SGML files with isolatin character-encoding.

CLEF (Cross-Language Evaluation Forum) Evaluation Package

ELRA-E0008 The CLEF Test Suite for the CLEF 2000-2003 Campaigns - Evaluation Package

The CLEF Test Suite contains the data used for the main tracks of the CLEF campaigns carried out from 2000 to 2003: Multilingual text retrieval, Bilingual text retrieval, Monolingual text retrieval, and Domain-specific text retrieval. It contains multilingual corpora in English, French, German, Italian, Spanish, Dutch, Swedish, Finnish, Russian, and Portuguese. The data consists of 1.62 Gb stored on 1 DVD.

The CLEF Test Suite is composed of:
- The multilingual document collections;
- A Step-by-Step documentation on how to perform a system evaluation (EN);
- Tools for results computation;
- Multilingual Sets of topics;
- Multilingual Sets of relevance assessments;
- Guidelines for participants (in English);
- Tables of the results obtained by the participants;
- Publications.