

**ECP-2007-LANG-617001**

**FLaReNet**

## **Existing evaluation and validation of LRs**

<b>Deliverable number/name</b>	<i>D5.1 – Existing evaluation and validation of LRs</i>
<b>Dissemination level</b>	<i>Public</i>
<b>Delivery date</b>	<i>31 May 2009</i>
<b>Status</b>	<i>Final</i>
<b>Version</b>	<i>1.0 (100131)</i>
<b>Author(s)</b>	<i>Jan Odijk, Antonio Toral</i>  <i>With contributions from:</i>  <i>Gilles Adda, Bogdan Babych, Luisa Bentivogli, Cristina Bosco, Tommaso Caselli, Anthony Hartley, Inma Hernaez, Valérie Mapelli and Djamel Mostefa</i>



### ***eContentplus***

This project is funded under the *eContentplus* programme<sup>1</sup>, a multi-annual Community programme to make digital content in Europe more accessible, usable and exploitable.

---

<sup>1</sup> OJ L 79, 24.3.2005, p. 1.

# 1 Table of Contents

- 1 TABLE OF CONTENTS ..... 2**
- 2 INTRODUCTION..... 3**
- 3 VALIDATION ..... 3**
  - 3.1 GENERAL INTRODUCTION ..... 3
  - 3.2 VALIDATION METHODOLOGIES ..... 4
  - 3.3 VALIDATION IN EUROPE ..... 6
  - 3.4 VALIDATION OUTSIDE OF EUROPE ..... 8
  - 3.5 VALIDATION RESOURCES ..... 9
  - 3.6 CHARACTERIZATION OF THE DEVELOPMENTS OF THE PAST 5 YEARS ..... 10
- 4 EVALUATION..... 11**
  - 4.1 GENERAL INTRODUCTION ..... 11
  - 4.2 EVALUATION METHODOLOGIES AND METRICS ..... 11
  - 4.3 EVALUATION IN EUROPE..... 12
  - 4.4 EVALUATION OUTSIDE OF EUROPE ..... 20
  - 4.5 EVALUATION PER CENTRE ..... 23
  - 4.6 CHARACTERIZATION OF THE DEVELOPMENTS OF THE PAST 5 YEARS ..... 25
- 5 BIBLIOGRAPHY ..... 28**
- A EVALUATION PACKAGES AVAILABLE AT ELRA..... 30**
- B LIST OF ACRONYMS AND PROJECT NAMES ..... 32**

## 2 Introduction

This document aims at describing the situation with regard to Validation and Evaluation of the past five years (so globally, 2004-2009). The aim of this document is to provide an analysis and charting of existing and widely accepted methodologies, facilities, campaigns, resources and services for validation and evaluation of LRs. Such an analysis will serve to identify those methods that best guarantee a useful assessment of their quality and characteristics, their maintenance requirements, and their efficacy for specific contexts of use. This will be the starting point for the next deliverable of this Work Package (D5.2: Validation and Evaluation Protocol for LRs: how to make reusable and maintainable quality LRs, including recommendations for maximizing re-usability of LRs).

Writing the present document started in a collaborative manner by the members of [FLaReNet's Working Group 5](#) (Evaluation Protocols and Procedures) under the coordination of the Working Group leader. In order to do so, we have used the wiki environment set up in the FLaReNet website. This has been the first activity of the project in which collaborative web technologies have been exploited in order to gather feedback from experts of the community. We can conclude that this first experience has been positive as we have received valuable contributions from several experts (see contributors in the author section of the deliverable, page 1). The downside is that result turned out to be a bit unbalanced; clearly contributions tackled the subject at different levels of depth and extension, resulting in some sections with a high level of detail whilst other are less detailed. The Working Group leader has taken the results of the collaborative effort and used it as a basis for making the current document in which this unbalance is corrected.

The deliverable is split up into two main sections: a section on Validation, and a section on Evaluation. Each main section is concluded by an assessment of the situation with regard to its topic. As will be clear, at least for some resources and components, some developments in the area of evaluation, especially the set-up of technical infrastructures for HLT technologies and their components may obviate the need of validation for some of these components.

## 3 Validation

### 3.1 General Introduction

Validation of a Language Resource (LR) is checking whether the LR is compliant with its specification and/or documentation, in short whether "you built the LR right". Validation can be *formal* ("is the form of the LR compatible with its specification/documentation?") and /or related to content, i.e. *substantive* ("is the content of the LR compatible with its specification/documentation?").

The term is used as defined here, though some researchers often use "validation" and "evaluation" interchangeably.

In principle, validation can be applied both to data and to tools/technologies/components. In practice, however, validation has been mainly applied to data, while for tools/technologies mainly evaluation has been applied.

Validation as an explicitly separate activity arose from a series of European cooperation projects, in particular the projects from the SpeechDat family of projects carried out in the period 1995-2005. Validation of resources was of utmost importance in these projects because

resources made by different project partners were exchanged among the project partners, which requires some confidence that all resources are equivalent in content and quality. Validation provided this confidence.

Validation as an explicitly separate activity was picked up by ELRA/ELDA, and elaborated into a specific methodology with validation manuals. Though validation was originally mainly applied to spoken resources, it was extended in a systematic manner by ELRA/ELDA to written resources. Validation in the cooperation projects usually consisted of a check of a resource against its specifications (or, in some cases, explicitly formulated validation criteria), but for resources produced in different contexts often only a check against the documentation is possible.

## **3.2 Validation methodologies**

### **3.2.1 Validation methodologies for Spoken Language Resources**

A series of projects where Spoken Language Resources (SLRs) were produced enabled to set up a number of validation criteria. These are in particular:

- [SpeechDat\(II\) project: "Validation criteria"](#) (pdf), Deliverable SD1.3.3, 5, November 1997, Henk van den Heuvel
- [SpeechDat\(E\) project: "Validation criteria"](#), Deliverable ED1.4.2, 27, October 1999, Henk van den Heuvel
- [SpeechDat-Car project: "Validation criteria"](#), Deliverable D1.3.1, 12, September 2000, Henk van den Heuvel
- [SpeeCon project: "Definition of Validation criteria" of SpeeCon](#) (pdf), Deliverable D41, 25 April 2002, Henk van den Heuvel, Shaunie Shammass, Ami Moyal, Oren Gedge
- [Orientel project: "Specification of Validation Criteria" of Orientel](#) (pdf), Deliverable D6.2, 25 August 2002, Dorota Iskra, Henk van den Heuvel, Oren Gedge, Sherrie Shammass

In addition, a validation manual was produced through the validation work carried out at ELRA in January 2000 and is also available online:

- [ELRA's Validation manual for SLR](#) (pdf), Validation of Content and Quality of Existing SLR: Overview and Methodology, Deliverable 1.1, 21 January 2000, Henk van den Heuvel, Louis Boves, Eric Sanders

### **3.2.2 Validation methodologies for Written Language Resources**

Inspired by other projects where written LRs were built and by the validation work carried out for SLRs, some methodologies were produced and were gathered by ELRA:

- [Validation of Linguistics Corpora](#), 28 April 1998, Tony McEnery, Lou Burnard, Andrew Wilson and Baker
- [Validation Manual for Lexica](#), Release 2.0, January 2004, Hanne Fersøe
- [Towards a Standard for the Creation of Lexica](#), May 2003, Monica Monachini, Francesca Bertagna, Nicoletta Calzolari, Nancy Underwood, Costanza Navarretta

- [EAGLES/ISLE](#) Meta Data Initiative web site

### 3.2.3 Full validation versus quick quality check

A full validation protocol includes a list of detailed elements that need to be checked. For instance, for SLRs, the typical elements are:

- Documentation: correctness & clarity
- Formats: directory structure & formats and names of files
- Design: completeness of recordings
- Speech files: quality in terms of clipping, SNR, etc.
- Lexicon: completeness & correctness of formats and transcriptions.
- Speakers: realistic distributions over gender, age, accents.
- Recording environments.
- Orthographical transcriptions: format & correctness.

Extensive validations are time-consuming and costly. ELRA and its Validation Committee (VCom) instructed SPEX (SPeech EXpertise centre, the Netherlands) to develop a method for a quick validation of a database. As a result, SPEX introduced the Quick Quality Check (QQC) for Speech LRs, whose methodology was also adapted to Written LRs as a second step.

As points of departure for the Quick Quality Check, two principles are taken:

1. The QQC mainly checks the database contents against its documentation. The main purpose of a QQC is to check if the documentation of the LR gives a correct account of the contents of the LR, in other words, if the LR meets the internal standards set up in the documentation.
2. Generally, the QQC of a LR should not take more than half a day's work.

The topics checked in a QQC are basically the same as those in the list of validation elements presented above. The crucial difference with a full validation is that a QQC only comprises a number of formal checks to see if the database contains what the documentation promises. There are no checks on the contents, that is the correctness of, say, for SLRs, orthographic and phonemic transcriptions. The report is concluded by a brief recommendation to the producer from the Validation Centres.

### 3.2.4 Validation by comparison and conversion

The validation of existing Natural Language Processing (NLP) models and resources strongly depends on the possibility of generalizing their results on data and languages other than those on which they have been trained and tested, i.e. usually English. A valuable contribute to the validation of existing models and data comes therefore from experiences that allow for consistent comparisons among approaches and representation schemes establishing shared standards, resources, tasks and evaluation practices with reference to various languages. For instance, in the [Evalita 2009 Parsing task](#), the contest among Italian parsers, the same training and testing data -from the [Turin University Treebank \(TUT\)](#)- are available both in

dependency-based and in constituency-based Penn-like annotation, thus allowing for a comparison between the two major existing parsing models. The data for dependency parsing is available both in ISST and TUT format, thus allowing for the comparison between two annotation schemes for the same language.

Moreover, the process of conversion of a LR from an annotation format *A* to an annotation format *B* is in itself a form of validation of the resource. The automatic conversion will in fact make explicit various limits (e.g. errors and inconsistent annotations) of the input. For example, within the TUT project for Italian, conversion tools have been applied to the native dependency-based treebank in order to generate a Penn-like annotation. As a side effect, two other formats that show intermediate layers of variation/similarity with respect to the TUT and Penn in terms of both richness of functional-syntactic information (i.e. amount and specificity of grammatical relations) and type of linguistic framework (i.e. constituency versus dependency, or minimal versus maximal projection) have been developed. This resulted in a strongly improved resource in terms of quality of the annotation.

### 3.3 *Validation in Europe*

#### 3.3.1 [Validation at ELRA](#)

In order to ensure a qualitative distribution, LRs must be subject to quality control and validation. The term *validation* in ELRA is used in reference to the activity of checking the adherence to standards, and the quality control of the LR product. The contribution of ELRA can be seen as a validation of existing and newly developed resources and documentation of the results in the catalogue.

To perform this task, a *validation committee*, VCom was set up by ELRA on 23rd October 2000. The aim of VCom is to maximise the "ease of use" and "suitability" of the LRs which may be needed for LE-systems, such as speech recognition, character recognition or information retrieval systems. For promoting "ease of use", VCom pushes forward the quality of LRs, i.e. by providing LRs with optimal documentation and minimal errors. For promoting "suitability", VCom supports standards and best practices for LRs leading to the best performance of state-of-the-art LE-systems.

At the beginning of VCom, ELRA established two Validation Centres (VCs) that are controlled and coordinated by VCom and the ELRA Board. The main tasks of the VCs are:

- Describe the quality of existing LRs.
- Improve the quality of existing LRs.
- Communicate with users and producers of LRs.
- Promote standards and best practices.
- Maintain the ELRA web pages concerning validation, according to the progress achieved within VCom.
- Maintain their LR validation portal.

Each VC is in charge of dealing with one area of activity, respectively:

- **Speech Language Resource Validation:** SLR validation concerns the quality evaluation of a database against a checklist of relevant criteria. These criteria are

typically the specifications of the databases, together with some tolerance margins for deviations. For this task [SPEX](#) was appointed (see section 3.3.2).

- Written Language Resources Validation: Aiming to fulfill its objectives regarding the production of validation manuals, the ELRA VC for WLR is working in close co-operation with highly recognised research centres in order to produce such manuals. The work being carried out capitalises on previous projects including, but not limited to, EAGLES, Parole, Simple, Multext, and previous work done by ELRA. For this task [CST](#) was appointed (see section 3.3.3). CST cooperates with a network of expert centres in Europe.

### 3.3.2 [Validation at SPEX](#)

The Speech Processing EXpertise centre (SPEX) in Nijmegen, the Netherlands, established a Validation Centre in 1999 in order to perform SLR validation. The centre aims to carry out the following tasks:

- Develop methodologies for SLR validation
- Carry out Quick Quality Checks on existing SLR in ELRA's catalogue
- Maintain a bug report service for SLR
- Assemble patch files for corrected SLR

### 3.3.3 [Validation at CST](#)

The Centre for Sprogteknologi (CST), Denmark, established a Validation Centre in 2002 in order to perform Written Language Resources (WLR) validation, encompassing both Lexica and Corpora. The centre has the following tasks:

- Develop methodologies for WLR validation and Quick Quality Checks
- Carry out full validations and Quick Quality Checks of existing WLR in ELRA's catalogue
- Quality standards for WLR

To ensure a broad enough range of expertise in both the different types of LRs and the languages they represent, a nucleus of a [network of experts](#) from different institutions who can participate in validations was set up.

### 3.3.4 [Validation at BAS](#)

The Bavarian Archive for Speech Signals (BAS), a public institution hosted by the University of Munich, performs validation of both internal and external speech and language resources. Validation is defined by BAS as the “formal check of a resource with respect to its specifications. This includes the checking of formats, documentations and structure, the completeness, the labeling, tagging etc”. BAS uses a standardised protocol for validation defined in the [BITS project](#) ([BAS](#) Infrastructures for [T](#)echnical [S](#)peech [P](#)rocessing).

### 3.3.5 [Validation at BASSS](#)

BAS Services Schiel (BASSS) is a spin-off originating in BAS (see section 3.3.4). BASSS focuses its business on the production, validation and evaluation of LRs mainly regarding speech corpora. The validation carried out regards:

- Documentation: complete and error free.
- Speech signal: quality checks, formal checks, completeness
- Annotations: sample checks, formal checks, correctness, completeness
- Metadata: formal checks, correctness, completeness

BASSS has also developed *WebTranscribe*, a web based tool for the manual validation of speech samples.

### 3.3.6 [Validation in the STEVIN programme](#)

STEVIN is a Flemish/Dutch cross-border Human Language Technology (HLT) Research and Development (R&D) programme that aims to contribute to the further progress of HLT R&D in Flanders and the Netherlands and stimulate innovation in this sector. In addition, it will strengthen the economic and cultural position of the Dutch language in the modern ICT-based society. The STEVIN-programme was launched in 2005 and will run until 2011 with a budget of 11.4 million Euros. It is jointly financed by the Flemish and Dutch governments and is coordinated by the Dutch Language Union.

In the programme several [types of projects](#) have been started up, including projects to create basic LRs (tools and data) for the Dutch language, projects to carry out fundamental research, projects to carry out application-oriented research, and a variety of demonstrator projects aimed at stimulating the HLT industrial sector. It has also organised itself or supported a variety of networking meetings for knowledge exchange and transfer, including promotional events and educational activities and projects.

Though Validation and Evaluation were not the primary focus of the [STEVIN programme](#), it does require that each project explicitly includes an evaluation component for technology, and a validation plan for basic LRs. Most of the basic LRs created have indeed been validated by independent parties (inter alia CST and BAS). In some cases this required extensions of the validation methodologies to resources of new types not validated before. One project in the STEVIN-programme, N-BEST, actually has evaluation of speech recognition systems as its main focus. It is discussed in section 4.3.4.

## 3.4 **Validation outside of Europe**

As sketched in the introduction, validation as an explicitly separate activity arose in Europe, and was extended there mainly by ELRA/ELDA and by the Bavarian Archive for Speech Signals (BAS). Outside of Europe, validation of resources has largely remained an implicit activity (as part of other activities to ensure quality) [Cieri 2006]. For the US, this difference may in part be ascribed to different resource production models, with the Linguistic Data Consortium (LDC) playing a more centralized role in the production of resources than ELRA/ELDA in Europe. This itself might be related to the different linguistic situation in Europe (many languages) vs. the US (few languages).

Validation also appears not to be an explicitly separate activity in Asia. We are not aware of individuals working on it, though some occasionally participate in relevant workshops.<sup>1</sup> It is illustrative that for example the SITEC (<http://www.sitec.or.kr/English/>) website, the GSK ([http://www.gsk.or.jp/index\\_e.html](http://www.gsk.or.jp/index_e.html)) website, and the NII-SRC website (<http://research.nii.ac.jp/src/eng/index.html>) do not contain an occurrence of the word “validation” and do not mention this term (or a related one) in their objectives.<sup>2</sup>

### **3.5 Validation Resources**

Validation resources can be divided into two types: formal validation resources and substantive validation resources.

#### **3.5.1 Formal validation resources**

Formal validation resources are tools which allow producers and users to validate the correctness of the LR with respect to the specification described in the documentation of the LR.

##### **3.5.1.1 XML Schema**

The use of XML schemes is now a standard for the formal validation of LRs. XML Schema was approved as a W3C Recommendation in 2001 (a second edition incorporating many errata was published in 2004). XML Schemas are a means to describe the structure, content and semantics of XML documents in detail. Different types of XML Schemas can be provided according to the source XML description.

#### **3.5.2 Substantial validation resources**

Substantial validation of LRs qualifies as an analysis of the elements which are present in the LR. In contrast to formal validation, substantial validation aims at verifying the correctness of the content of the elements which compose the LR according to the specifications described in the documentation. Substantial validation is mainly conducted manually and on reduced portions of the LR and in some cases with the help of graphical tools (for instance VALIDO, Navigli 2006). As far as automatic substantial validation is concerned, it is possible to identify different methodologies:

- use of *n*-gram methods (e.g. van Noord 2004, Sagot & de la Clergerie 2006, de Kok et al. 2009);
- comparison of output of different tools;
- graph-based techniques.
- frequency counts

---

<sup>1</sup> For example, Chu-Ren Huang and Takenobu Tokunaga in the LREC 2006 *Workshop on Quality Assurance and Quality Measurement for Language and Speech Resources*.

<sup>2</sup> At least, this holds for the English versions of these websites.

### **3.6 Characterization of the Developments of the past 5 years**

Validation as an explicitly separate activity originated in the SpeechDat Family of projects in Europe, was taken up and further actively promoted by ELRA/ELDA, and also played an important role in the creation of resources in projects in the Dutch-Flemish STEVIN programme. In the context of these projects and programmes a number of concrete validation manuals and guidelines were developed and applied illustrating the currently best practices with regard to validation.

However, validation as an explicitly separate activity has never extended beyond Europe, and appears to be losing ground in Europe as well. If validation of a resource, which always costs some effort and money, is not budgeted explicitly, for example by being enforced by funding agencies or cooperating partners, it usually will not be done. Distribution agencies sometimes enforce validation of resources, but usually their budgets are limited so that only a minimal validation ('quick quality check') is performed.

Though some advocate that validation can be carried out by comparison and conversion, this is true only to some extent: as long as the comparison and conversion is a human effort, human intelligence will compensate for any deviations between documentation and actual resource, so the danger exists that such deviations are never made explicit or adapted. The situation is different if these comparisons and conversions are done in a fully automated manner, and done in accordance with requirements that are independent of the specific resource being validated.

At least in part, validation may be replaced by the requirement for interoperability of resources and tools, as for example in the CLARIN infrastructure. Henk van den Heuvel (2009) pointed out that standardization and the definition of appropriate metadata sets will allow automated validation (since metadata can be seen as formalized parts of the documentation). Interoperability will require even stronger requirements, and will be an excellent test for both formal and substantive compliance with the documentation.

Automation and evaluation of the tools doing the automation will also contribute to making validation easier and cheaper (see van den Heuvel (2009) and Schiel (2009)), but it is not obvious that this can also be put to use for completely new types of resources that are developed to investigate new research questions or to develop new technologies.

## 4 Evaluation

### 4.1 General Introduction

Evaluation of an LR (broadly construed as including technologies and applications) is checking whether the LR is suited for a specific task, in short whether "you built the right LR". Though evaluation can in principle be applied both to data and to technologies//tools, in actual practice evaluation is mostly applied to technologies and applications only. This section discusses different methodologies used for evaluation during the last years, reports on activities related to evaluation (projects, evaluation campaigns, centers, etc) both in Europe and outside Europe and finally tries to extract the general lines that have characterised the research in this area during this period. For a different but rather complete and reasonably up-to-date overview of evaluation, see the [ELRA HLT Evaluation Portal](#).

### 4.2 Evaluation Methodologies and Metrics

Multiple views on evaluation of technologies are around. We focus here on evaluation of technologies in isolation.

The trend of the past years has been to use methodologies and metrics for automating evaluation. Non-automated evaluation requires a lot of (human) effort and time, and is therefore quite expensive. Research, however, requires frequently repeated evaluation experiments. Automating evaluation, if it can be done in a reliable manner, can solve these conflicting requirements.

Evaluation always requires creating a reference, and here expensive and time-consuming human effort usually cannot be avoided. Automating evaluation experiments is relatively easy and common when a single reference can be defined and formally represented. For example, in the area of speech recognition, one usually defines a single reference transcription of the speech, which makes it possible to make a relatively simple measurement of the results of an evaluation experiments, for example using Word Error Rate (WER) as metrics. In contrast, for evaluating the quality of speech synthesis it is very difficult to define and formally encode a reference, which makes it necessary to do evaluation using judgments by (multiple) humans (usually by Mean Opinion Score (MOS) experiments).

For the development of many NLP-technologies several metrics are used, e.g. accuracy and metrics derived from it. The metrics of Recall, Precision, and F-Score, originally used almost exclusively in information retrieval (Rijsbergen 1979) have gained increasingly more ground.

In the domain of machine translation, each new evaluation experiment originally required human effort since it is usually very difficult to define a single reference (since there are multiple correct translations, which may differ in lexical selection and syntactic constructions, different distribution over sentences, etc.), but in this domain several proposals were made to automate this. However, it clearly illustrates the difficulty of selecting an evaluation metrics that allows automation and provides reliable results.

## 4.2.1 Evaluation Metrics in MT

The trend to automate evaluation in the area of MT was initiated by seminal work by Papineni, K. *et al.* (2002). They introduced BLEU as an evaluation metrics. BLEU makes it possible to automatically measure (aspects of) translation quality of a candidate translation against a reference corpus which contains multiple reference translations, and it is claimed that it yields results that correlate highly with human judgments of quality at the corpus level.

The NIST metric is derived from the BLEU metric, with assignment of more weight to n-grams that are rarer (Doddington 2002).

Position-independent word error rate (PER), is based on the Word Error Rate, familiar from speech recognition evaluation. PER, in contrast to WER, allows for re-ordering of words and sequences of words between a translated text and a reference translation.

The Translation Error Rate (TER) metrics measures the number of edits needed to change a system output so that it exactly matches a given reference. It differs from WER in that it counts a shift of any sequence of words over any distance as a single edit; hence this has the same costs as an insertion, deletion or substitution (Snover 2006).

The METEOR metric is designed to address some of the deficiencies inherent in the BLEU metric which focuses on precision. METEOR also includes recall aspects, by using the weighted harmonic mean of unigram precision and unigram recall (Lavie 2004). It is claimed to achieve a higher correlation than BLEU and NIST which are based on precision alone. METEOR also includes a mechanism for synonym matching so that also synonyms for words yield a match.

Many other metrics, usually small variants of the metrics mentioned above, have been proposed and are being tested.

Evaluation metrics inspired by BLEU have also been used for domains other than MT, e.g. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and its variants for text summarisation (Lin 2004).

Automated MT evaluation tools, such as BLEU or NIST, initially proposed for assessing the progress in the development of individual MT systems, have also been used in other evaluation scenarios, e.g., for comparing translation quality of different MT systems, or for assessing the difficulty of specific genres and text types for MT. (Babych *et al.* 2004), for predicting human evaluation scores and acceptability thresholds for MT systems on the basis of automated evaluation scores. However, some of such scenarios were later shown to be problematic. For instance, it is not possible to make a meaningful comparison of Rule-Based vs. Statistical MT systems using BLEU, because this metric over-rates the quality of SMT (Callison-Burch *et al.* 2006). Also, absolute values of BLEU-type scores are not meaningful without comparison with the values obtained under the same experimental conditions, because they cannot predict human evaluation scores without setting specific calibration parameters (e.g., slope and the intercept of the regression line), but these parameters are different for different text types and target languages (Babych *et al.* 2005). Therefore, MT systems cannot be meaningfully assigned BLEU scores for comparison with other MT systems across the board, and need to be evaluated on the case-by-case basis under controlled experimental conditions. As a result, the validity of many automated MT evaluation metrics for industrial use been questioned (Thurmair 2007).

## 4.3 Evaluation in Europe

In Europe, several projects and programmes carried out evaluation, either as their main activity (e.g. CLEF), or as an important part of a more encompassing work package (e.g. ECESS). The most important ones are described here.

### 4.3.1 [CLEF](#)

CLEF (the Cross Lingual Evaluation Forum) organises annual technology evaluation on Information Retrieval since 2000. Since 2008 CLEF is included in the activities of the [TrebleCLEF Coordination Action](#) funded by the Seventh Framework Program of the European Commission. The aim of CLEF is to promote research in the field of multilingual systems by organizing evaluation campaigns in which a series of tracks are developed to test aspects of mono- and cross-lingual information retrieval. Ten evaluation campaigns have been organized so far (see also Peters 2008, 2009b).

- Ad Hoc (Multilingual Textual Document Retrieval): Development and evaluation of monolingual and cross-lingual textual document retrieval systems. In the 2009 campaign the focus has been enlarged to include both European languages and Persian. The document collections include newspaper and news agency documents and library catalog records (provided by The European Library – TEL). The track offers three tasks: Tel@CLEF, Persian@CLEF and Robust-WSD;
- iCLEF (Interactive Cross-Language Retrieval): Evaluation of cross-language information extraction based on user-perspective;
- QA@CLEF (Multilingual Question Answering): Evaluation of capabilities of systems on question answering. The 2009 campaign three tasks have been proposed:
  - ResPubliQA: evaluation on question answering systems in the law domain from the European legislation (JRC-Acquis collection of EU parliamentary documents). In this task the systems are evaluated on the retrieval of the passage which answers the question and not on the exact question;
  - QAST: evaluation of multilingual question answering systems in speech scenario in which written and oral questions are formulated against a set of transcribed audio recordings
  - GikiCLEF: evaluation of open domain questions over Wikipedia which require geographical reasoning, complex information extraction and cross-lingual processing. The document collection included Bulgarian, Dutch, English, German, Italian, Norwegian, Portuguese, Romanian and Spanish
- ImageCLEF: Evaluation on the retrieval from visual collections employing both text and image retrieval techniques;
- VideoCLEF (Cross-language Video Retrieval): Evaluation on video content in a multilingual environment;
- CLEF-IP (Intellectual Property): This task has been first introduced in the 2009 campaign. Evaluation of cross-lingual information access on patent documents;
- LogCLEF: Evaluation of queries and other logged activities in order to understand search behavior in multilingual contexts;
- Grid@CLEF (Grid Experiments): This evaluation track has been introduced as a pilot experiment in the 2009 campaign. Participants have to conduct experiments according to the CIRCO (Coordinated Information Retrieval Components Orchestration) protocol.

The CLEF (TrebleCLEF) projects also yielded descriptions of best practices in the area of information retrieval evaluation (e.g. Braschler & Gonzalo 2009, Moreau 2009, Sanderson & Braschler 2009)

### 4.3.2 [EVALDA](#)

EVALDA is a French evaluation project of HLT technologies for French launched in 2003. This project is financed by the French Ministry of Research in the context of its Technolanguage programme. The aim of the project is to establish a permanent evaluation infrastructure for the language engineering sector in France and for the French language. Six evaluation campaigns have been organised:

- [ARCADE II](#) : Evaluation of bilingual text and vocabulary alignment systems. Following the success of ARCADE I, this follow-up campaign aimed to evaluate alignments between more distant or 'exotic' languages i.e. Greek, Russian, Japanese, Chinese.
- [CESART](#) : Evaluation of terminology extraction tools, including tools for extracting ontologies and semantic relations.
- [CESTA](#) : Evaluation of Machine Translation Systems. French was the pivotal language, however, several languages from and into French were used (English, Spanish, German, and Arabic).
- [EASY](#) : An evaluation campaign designed to test syntactic parsers. A side effect of the campaign is the creation of a syntactically parsed reference text composed of several genres of text (newspapers, literary texts, electronic texts etc).
- [EQUER](#) : Evaluation of Question/Answering systems. Three reference corpora were used: a large general corpus (newspapers, general texts), a web corpus and a corpus made up of medical texts.
- [ESTER](#) : Evaluation of automatic broadcast news transcriptions systems. This campaign included the evaluation of segmentation tasks and identification of named entities.
- [EVASY](#) : Evaluation of Speech synthesis systems. This campaign was to feature a novel method for the evaluation of prosody in synthesized speech.
- [MEDIA](#) : Evaluation of Man-Machine dialogue systems. In this case, the task of hotel room reservation (including some local tourist information) is envisaged.

See also [http://www.technolanguage.net/article.php3?id\\_article=331](http://www.technolanguage.net/article.php3?id_article=331).

### 4.3.3 [EVALITA](#)

EVALITA is an evaluation campaign of NLP tools for Italian. So far, two editions have been organised (2007 and 2009). It covers topics such as Part-of-Speech tagging, parsing, Named Entity Recognition, Word Sense Disambiguation and Lexical Substitution. Besides the traditional tasks proposed in EVALITA 2007, the 2009 campaign has two important novelties. In conjunction with AISV (Italian Association of Speech Science) EVALITA 2009 also organises the first speech input technology evaluation for Italian with tasks such as Speaker Identity Verification. Moreover, for exploring possible links with Parsing evaluation for other languages, the EVALITA campaign has a tiny development and test corpus shared with [PASSAGE-2](#), the French campaign on dependency parsing. Aligned data both in French and Italian have been hand-annotated (200 sentences of development and 50 sentences for test) both with PASSAGE annotations for the French part and the TUT annotations for the Italian one.

The general objective of EVALITA, which is supported by the NLP working group of [AI\\*IA](#) (the Italian Association for Artificial Intelligence), is to promote the development of language and speech technologies for the Italian language, providing a shared framework

where different systems and approaches can be evaluated in a consistent manner.

#### 4.3.4 [N-BEST](#)

The N-BEST project has been developed under the STEVIN programme (section 3.3.6). The goal of the N-BEST project is to evaluate the performance of present-day large vocabulary continuous speech recognition (LVCSR) systems for Dutch. N-BEST qualifies as the first attempt in this field for Dutch. The task is to identify all spoken words in a set of given excerpts of audio (automatic speech identification, ASR, or speech-to-text, STT). So far only one campaign has been conducted in 2008.

#### 4.3.5 [ESTER2](#)

The ESTER 2 campaign was held in 2008 and 2009. The campaign is organised by the French-speaking Speech Communication Association (AFCP French-speaking ISCA Regional Branch) and the French Defense expertise and test center for speech and language processing (DGA/CEP) in collaboration with ELDA. The ESTER 2 campaign is a continuation of the ESTER campaign (2003-2005). The ESTER 2 campaign aims at evaluating systems for automatic speech identification for French. With respect to the first campaign, the ESTER 2 campaign has extended the variety of speaking styles, genres and accents by including broadcast news, entertainment shows and debates and shows from an African French-speaking radio. The campaign implemented three tasks: segmentation, transcription and information extraction. The core tasks of the campaign were Sound Event Tracking (SED), Speaker Tacking (SVL), speaker Diarization (SRL), Transcription (TRS) and Named Entity Detection (NE).

#### 4.3.6 [Passage](#)

The PASSAGE project (ANR-06-MDCA-013) organises an evaluation campaign for parsing as a continuation to the EASY campaign of the EVALDA project (see section 4.3.2). The aims of PASSAGE are:

- The creation of a tree-structured corpus which will be available to the research community.
- Exploration of improving possibilities of parsing systems by lexical acquisition derived from the combination of different analysis.

PASSAGE should then improve the accuracy and robustness of existing French parsers and exploit the resulting syntactic annotations to create richer and more extensive linguistic resources.

#### 4.3.7 [InFile](#)

INFILE (INformation, Filtering, Evaluation) is a cross-language adaptive filtering evaluation campaign organised by the [CEA LIST](#), [ELDA](#), and the [University of Lille 3](#). The goal of INFILE is “to organize evaluation campaigns for monolingual and multilingual information filtering systems based on close-to-real-usage conditions for intelligence applications”. INFILE 2009 is a pilot task in CLEF (see section 4.3.1) and is scientifically endorsed by

TREC (see 4.4.2). It proposes two tasks (interactive filtering and batch filtering) for three languages (Arabic, English and French).

#### 4.3.8 [TC-STAR](#)

The EU FP6 TC-STAR project (2004-2007), was envisaged as a long-term effort to advance research in the core technologies of Speech-to-Speech Translation (SST). To assess the advances in SST technologies, annual competitive evaluations were organised (e.g. [2006](#), [2007](#)). The aim of the evaluation campaigns were to measure the progress made during the life of the project in Automatic Speech Recognition (ASR), Spoken Language Translation (TTS), Text-To-Speech and in the whole end-to-end Speech-to-Speech system. In addition to the measure performance, the infrastructure built in TC-STAR was also evaluated. For Automatic Speech Recognition, systems were evaluated automatically for English, Spanish and Mandarin Chinese by computing Word Error Rates and Character Error Rates. For SLT, evaluation carried out for English-to-Spanish, Spanish-to-English and Mandarin Chinese-to-English. In addition to automatic metrics such as BLEU, NIST, WER, PER, subjective evaluations were organised with hundreds of evaluators to assess the quality of SLT systems. For TTS, systems were evaluated for Chinese, Spanish and English. MOS subjective tests were organised to assess various aspects such as *overall voice quality*, *listening effort*, *comprehension*, *pronunciation*, *naturalness*, etc. In addition to subjective tests, individual modules evaluations were carried out.

#### 4.3.9 [EUROMATRIX](#) and [EUROMATRIXPlus](#)

EuroMatrix (running from 2006 to 2009, funded by the EU under FP7) aims at a major push in Machine Translation (MT) technology by applying the most advanced MT technologies systematically to all pairs of EU languages with special attention for the languages of the new and near-term prospective member states. It designs and investigates novel combinations of statistical techniques and linguistic knowledge sources as well as hybrid MT architectures.

EuroMatrix aims at enriching the statistical MT approach with novel learning paradigms and experiment with new combinations of methods and resources from statistical MT, rule-based MT, shallow language processing and computational lexicography/morphology. With respect to evaluation, EUROMATRIX aims to organize a competitive annual international evaluation of MT with a strong focus on European economic and social needs

EuromatrixPlus is the successor of EuroMatrix, will run from 2009 to 2012 and is funded by the EU under FP7. Regarding evaluation, the project aims to

- Organize an annual evaluation campaign on European language translation on large-scale open domain tasks such as translation of news stories or Wikipedia articles.
- Prepare ready-for-use training and test sets, and annotate them with additional linguistic markup.
- Develop manual and automatic evaluation metrics and validate these metrics.
- Pose special challenges that arise from work on the user-centric work packages, for instance improved interactive machine translation.

#### 4.3.10 PORT-MEDIA

PORT-MEDIA is a follow-up of the MEDIA campaign of the EVALDA project (see section 4.3.2). The aim of the project is to provide the MEDIA corpus with three additional aspects of great importance in spoken dialog systems: robustness, portability across domains and languages and rich structures for high-level semantic knowledge representation. An important outcome of the project will be the availability of a platform for the evaluation of automatic speech understanding systems with manual or automatic transcriptions.

#### 4.3.11 ECESS

ECESS (European Center of Excellence for Speech Synthesis) is an open, non funded consortium for institutions active in speech synthesis and related topics. ECESS is targeted to build an infrastructure with the goal to accelerate progress in speech synthesis with respect to models, algorithms, and languages. The infrastructure is based on the idea to facilitate the exchange of modules, language resources, and tools needed for speech synthesis and related topics. The exchange is based on the principles of validation and evaluation.

The major goals of ECESS are

- Achieve the critical mass needed to push substantially progress in speech synthesis technology
- Integrate basic research know-how related to speech synthesis
- Attract public and private funding

ECESS is building an infrastructure allowing for each institution to be active for a specific research task and to benefit from the activities of the other institutions. Thus ECESS acts as a large 'virtual' institute, which has the critical mass for fastening progress in speech synthesis. The basic elements of the infrastructure of ECESS are:

- Common system architecture based on well defined modules and interfaces,
- Common set of specified language resources,
- Common set of tools,
- Common set of evaluation criteria defining the quality of modules, language resources and tools.

The functionality and interfaces of the ECESS TTS modules and the procedures for evaluation have been developed by a joint effort between ECESS and the EC-funded project TC-STAR. Within TC-STAR three evaluation campaigns on speech synthesis have been performed conducted by ELDA. TC-STAR ended in 2007, but ECESS aims to continue the approach initiated there.

Although the evaluation carried out in ECESS is specific to the area of speech synthesis, it includes a lot of other aspects that make it an attractive model for other technologies as well.

- A set of components and their interfaces for speech synthesis systems is defined and agreed upon by the partners
- If a partner contributes two components it gets access to the components contributed by other partners
- Evaluation can be done not only on whole systems but also on components of the whole system
- Sharing components avoids duplication, allows research groups to focus on their speciality, creates *de facto* standards and in fact even actual interoperability!

- The consortium includes commercial parties and rights and obligation are secured by a clear consortium agreement.
- It incorporates some good ideas that were launched and implemented in the SpeechDat family of projects but adapts these to a new domain of (partially shared) technology development and evaluation

#### 4.3.12 [IWSLT](#)

The International Workshop on Spoken Language Translation (IWSLT) is a campaign focused on the evaluation of spoken language translation technologies. The impact of spontaneity aspects on the ASR and MT systems performance as well as the robustness of state-of-the-art MT engines towards speech recognition errors are investigated in detail. Workshops have been held since 2004.

#### 4.3.13 [ALBAYZÍN](#)

Albayzín is an evaluation campaign on speech technologies performed by the [Spanish Thematic Network on Speech Technologies](#). This campaign is organized every two years, and the results and conclusions are presented in a special session that takes place at the "Speech Technologies Workshop". The campaign aims at promoting research in Speech Technologies, attracting young researchers to the field and reinforcing the links and relationship between researchers. The campaign is open to everybody and participation of groups that are not members of the Spanish Network is encouraged.

Each campaign focuses in a different area (Speech synthesis, speech/speaker recognition, translation...). Two campaigns have been organized so far, with the following topics:

- Albayzín 06: Speech Recognition of pathological voices, Imitation of voices to break the access on a voice controlled access system, Speaker segmentation and identification and Translation from text to sign-language.
- Albayzín 08: Language verification, Speech synthesis (Spanish), Text-to-text Translation (Spanish to Basque)

#### 4.3.14 [PASCAL/PASCAL2 Challenges](#)

PASCAL/PASCAL2 (Pattern Analysis, Statistical Modelling and Computational Learning) is a Network of Excellence funded by the European Union, started under the 6th Framework Program. PASCAL/PASCAL2 is developing the expertise and scientific results that will help create new technologies such as intelligent interfaces and adaptive cognitive systems. Under this program a total of 26 challenges have been created. We will report below only those evaluation campaigns which are directly related to LRs.

- [Morpho Challenge](#): Morpho Challenge is an evaluation campaign held every two years. The first edition was in 2005. The aim is to evaluate the performance of machine learning algorithm at discovering which are the morphemes (smallest individually meaningful units of language) which compose a word.
- [Consonant Challenge](#): Evaluation of ASR systems on the basis of human-computer comparisons on a task involving consonant identification in noise;

- [PASCAL Recognizing Textual Entailment \(RTE\)](#): The RTE Challenges have promoted research in textual entailment recognition as a generic task that captures major semantic inference needs across many natural language processing applications. Moreover, after the first three highly successful PASCAL RTE Challenges, RTE became a track at the NIST 2008 Text Analysis Conference (TAC) (see section 4.4.4). Up to now, the [RTE-5](#) campaign is underway, and particularly relevant to the issue of LRs evaluation is the introduction of ablation tests as a requirement for systems participating in the task. Ablation tests are required in order to collect data to better understand the impact of the knowledge resources used by RTE systems and evaluate the contribution of each resource to systems' performance. This experiment is very important as it represents a first step towards the definition of a new pilot task focused on knowledge resource evaluation, to be proposed in the RTE-6 campaign.
- [Letter-to-Phoneme Conversion Challenge](#): Evaluation of systems for speech synthesis, where input text has to be transformed to a representation that can drive the synthesis hardware, and necessary for some aspects of speech recognition. Letter-to-sound conversion qualifies as a class of problems, which include not just automatic pronunciation but stress assignment, letter-phoneme alignment, syllabification and/or morphemic decomposition, and so on. Both rule-based and ML systems participated.
- [PASCAL Ontology Learning Challenge](#): Evaluates the automated construction and population of ontologies. The tasks present several subtasks: ontology construction, ontology population, ontology extension and concept naming. The evaluation is carried out against a *gold* standard built by human annotators.
- [Assessing ML methodologies to extract implicit relations from documents Challenge](#): Evaluates the application of different machine learning algorithm to perform Information Extraction.
- [Large Scale Hierarchical Text Classification](#): Evaluates the classification of textual documents to the categories of a hierarchy. There are [four tasks](#) depending on the kind of information participants are allow to train their systems with.

#### 4.3.15 [FEMTI](#)

FEMTI, a Framework for Machine Translation Evaluation within the ISLE initiative (International Standards for Language Engineering), is “an attempt to organise the various methods that are used to evaluate MT systems, and to relate them to the purpose and context of the systems”. In order to do this, FEMTI provides [two interrelated classifications or taxonomies](#). The first classification enables evaluators to define an intended context of use for the MT system to evaluate. Each feature is then linked to relevant quality characteristics and metrics, defined in the second classification.

The aim of FEMTI is to help two types of users:

- Final users of MT systems. They can select the quality characteristics that are most important to them and thereby choose the MT system that best suits these characteristics.
- Designers and developers of MT systems. They can browse and select the characteristics that best reflect their circumstances, and thereby find associated evaluation measures and tests. They can also learn about the needs of users and find niche applications for their system.

## 4.4 Evaluation Outside of Europe

### 4.4.1 [GALE](#) (US)

The DARPA GALE (Global Autonomous Language Exploitation) program aims to develop systems for analysing and interpreting huge volumes of speech and text in multiple languages in order to convert and distil the data in easy-to-understand forms to military personnel and monolingual English-speaking analysts in response to direct or implicit requests. Under the DARPA GALE program evaluation campaigns have been developed in order to assess the quality of MT (Arabic/Chinese to English). The input data are in the form of either audio or text, with the output always being text. With respect to other MT evaluations the DARPA GALE program evaluates the quality of the translations by measuring the edit distance between a system output and a gold standard reference. The term "edit-distance" refers to the number of edits (modifications) that someone needs to make to the output of a machine translation system such that the resulting text is fluent English and completely captures the meaning of the gold standard reference.

### 4.4.2 [TREC](#) (US),

The Text REtrieval Conference (TREC) started in 1992 as part of the TIPSTER Text program and is an evaluation initiative co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense. The aim of TREC is to support research within the information retrieval community. The evaluation effort grows both in the number of participating systems and in the number of tasks every year. TREC has sponsored the first large-scale evaluations of the retrieval of non-English (Spanish and Chinese) documents, retrieval of recordings of speech, and retrieval across multiple languages. TREC has also introduced evaluations for open-domain question answering and content-based retrieval of digital video. The TREC test collections are large enough so that they realistically model operational settings. The 2009 tracks are reported below. For information of the past TREC tracks see <http://trec.nist.gov/tracks.html> .

- **Blog Track:** Evaluation of information seeking behavior in the blogosphere.
- **[Chemical IR Track](#):** Evaluation of systems and for large scale search in chemical documents including academic papers and patents.
- **[Entity Track](#):** Evaluation of systems for detecting entity-related search (e.g. entities and properties of entities) on Web data.
- **[Legal Track](#):** Evaluation of Information Retrieval systems in order to meet the needs of the legal community
- **[Million Query Track](#):** Evaluation of Information Retrieval systems by using large numbers of queries incompletely, rather than a small number more completely.
- **[Relevance Feedback Track](#):** Relevance feedback (RF) can be defined as the judging phase of a user on the relevance of existing information returned from a search on the basis of which the retrieval system performs a new search, returning more information to the user. This track aims at evaluating the effects of different factors on the success of relevance feedback.
- **Web Track:** Evaluation of systems performing Web-specific retrieval tasks, including diversity and efficiency tasks, over collections of up to 1 billion Web pages.

### 4.4.3 [Document Understanding Conferences \(DUC\)](#)

The Document Understanding Conferences (DUC) grew out of the TIDES (Translingual Information Detection Extraction and Summarization) programme and worked on continuing evaluation in the area of text summarization. It was sponsored by the Advanced Research and Development Activity ([ARDA](#)). The conference series were run by the National Institute of Standards and Technology ([NIST](#)) to further progress in summarization and enable researchers to participate in large-scale experiments. DUC has now been incorporated in TAC (section 4.4.4).

### 4.4.4 [ACE](#) / [TAC](#) (US)

The ACE (Automatic Content Extraction) program constitutes the NIST series of information extraction technology evaluation. The objective of this program is to develop automatic content extraction technology to support automatic processing of human language in text form from a variety of sources (such as newswire, broadcast conversation, and web logs). There have been nine evaluation campaigns from 1999 to 2008. Tasks covered relate to the detection and characterization of Entities, Relations, and Events. ACE is becoming a track in the Text Analysis Conference (TAC) in 2009.

TAC is a series of workshops that provides the infrastructure for large-scale evaluation of NLP technology. TAC's mission is to support research within the Natural Language Processing community by providing the infrastructure necessary for large-scale evaluation of NLP methodologies. TAC's primary purpose is not competitive benchmarking; the emphasis is on advancing the state of the art through evaluation results. In particular, the TAC workshop series has the following goals:

- Promote research in NLP based on large common test collections;
- Improve evaluation methodologies and measures for NLP;
- Build a series of test collections that evolve to anticipate the evaluation needs of modern NLP systems;
- Increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- Speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in NLP methodologies on real-world problems.

The tasks covered in the 2009 editions are Knowledge Base Population, Textual Entailment and Summarization.

### 4.4.5 [NTCIR](#) (Japan)

NTCIR (NII Test Collection for IR Systems) is a Japanese series of evaluation workshops designed to enhance research in Information Access technologies including Information Retrieval, Question Answering, summarization, extraction, etc. There have been 8 workshops since 1998. The aims of NTCIR are to:

- Encourage research in Information Access technologies by providing large-scale test collections reusable for experiments and a common evaluation infrastructure allowing cross-system comparisons

- Provide a forum for research groups interested in cross-system comparison and exchanging research ideas in an informal atmosphere
- Investigate evaluation methods of Information Access techniques and methods for constructing a large-scale data set reusable for experiments.

#### 4.4.6 [Senseval](#) / [Semeval](#)

The first three workshops, Senseval-1 through Senseval-3, were focused on word sense disambiguation, each time growing in the number of languages offered in the tasks and in the number of participating teams. In the 4th workshop, [SemEval-2007](#), the nature of the tasks evolved to include semantic analysis tasks outside of word sense disambiguation, such as classification of semantic relations between nominals, multilingual tasks, tasks around affective text, Time-Event temporal relations and others.

SemEval-2010 proposes a series of tasks aimed at the evaluation of semantic analysis systems. Examples include coreference, VP ellipsis detection and resolution, cross-lingual tasks, argument selection and coercion, noun compound interpretation, linking events and their participants in discourse, disambiguation of sentiment ambiguous adjectives, and others. The 5th workshop on semantic evaluation will take place mid 2010. There are many computer systems that do automatic semantic analysis of text. The purpose of SemEval is to evaluate the strengths and weaknesses of such systems with respect to different words, relations, types of texts, different varieties of language, and different languages.

#### 4.4.7 [Festvox Blizzard Challenge](#) (USA, CMU)

The Blizzard Challenge (Black and Tokuda 2005) is a multi-site challenge to build corpus-based speech synthesizers from a common database resulting in a synthetic voice that will be evaluated by a large number of listeners. There have been five annual editions the first taking place in 2005. The aim is to better understand different speech synthesis techniques on a common dataset, and to be able to compare different research techniques in building corpus-based speech synthesizers.

#### 4.4.8 [Recognizing Textual Entailment \(RTE\) Challenge](#)

The RTE Challenges have promoted research in textual entailment recognition as a generic task that captures major semantic inference needs across many natural language processing applications. Moreover, after the first three highly successful PASCAL RTE Challenges, RTE became a track at the NIST 2008 Text Analysis Conference (TAC). Up to now, the RTE-5 campaign is underway, and particularly relevant to the issue of LRs evaluation is the introduction of ablation tests as a requirement for systems participating in the task. Ablation tests are required in order to collect data to better understand the impact of the knowledge resources used by RTE systems and evaluate the contribution of each resource to systems' performance. This experiment is very important as it represents a first step towards the definition of a new pilot task focused on knowledge resource evaluation, to be proposed in the RTE-6 campaign.

## 4.5 *Evaluation per Centre*

### 4.5.1 ELRA/ELDA

ELRA (European Language Resources Association) and its operational body ELDA (Evaluations and Language Distribution Agency) build on extensive experience in national and European HLT projects in playing their role as major contributors to the creation of a European infrastructure for HLT evaluation. Long term involvement in the field helps ELRA and ELDA to provide a forum for evaluation activities as well as a warehouse of software packages and resources for various kinds of evaluation. ELRA's over-riding aim is to facilitate communication between the various stakeholders in the provision of adequate evaluation: funding agencies, the scientific academic community, the research and development communities in industry and the user community wishing to profit from accurate evaluation of human language technology products, as well as all other interested parties.

#### **Evaluation Resources**

Almost all evaluation projects in which ELDA took part ended up with an evaluation package.

Evaluation packages are publicly available through [ELRA's catalog of LRs](#). An evaluation package consists of full documentation (including definition and description of the evaluation methodologies, protocols, and metrics), along with the data sets and software scoring tools, necessary to evaluate developed systems for a given technology. Such a package therefore enables external participants to benchmark their systems and compare results to those obtained during the official evaluation campaign. Appendix A contains a list of packages currently available via ELRA.

In 2003, ELDA launched EVALDA, a French evaluation project of HLT technologies for French. Between 2004 and 2007 ELDA organized several [TC-STAR evaluation campaigns](#) and evaluation workshops ([2006](#), [2007](#)).

The project [CHIL](#) (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 1<sup>st</sup>, 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 includes 15 leading research laboratories from 9 countries, representing today's state of the art in multimodal and perceptual user interface technologies in European Union and the US. ELDA coordinated the data collection and technology evaluation of the project. Four evaluation campaigns were organized between 2004 and 2007 and for the following technologies were evaluated:

- Close-Talking Automatic Speech Recognition
- Far-Field Automatic Speech Recognition
- Speech Activity Detection
- Acoustic Event Detection
- Acoustic Environment Classification and Recognition
- Face and head tracking
- Head pose estimation
- Person tracking (acoustic, video and multimodal)
- Person recognition (acoustic, video and multimodal)

- Question answering
- Summarization

From 2000 until now, ELDA is involved in [CLEF](#) (see section 4.3.1). ELDA co-organised and co-organises several tracks such as Question Answering, AdHoc, Information filtering, etc. ELDA is also involved in the [TrebleCLEF](#) coordination action which promotes and support evaluation activities in Multilingual Information Access.

ELDA is involved in the FP7 project MEDAR, which addresses International Cooperation between the EU and the Mediterranean region on Speech and Language Technologies for Arabic. Within the project, evaluation of HLT for Arabic is organized by ELDA with a special focus on Machine Translation and Multilingual Information Retrieval.

ELDA is also organizing cross-language adaptive filtering evaluation campaigns through the [INFILE](#) project. The goal of the INFILE project is to organize evaluation campaigns for monolingual and multilingual information filtering systems based on close-to-real-usage conditions for intelligence applications. Both methodology and metrics are discussed within a group of experts, set up at the beginning of the project. Two evaluation campaigns were carried out in 2008 and 2009.

The French project PASSAGE motivations are to improve the accuracy and robustness of existing French parsers and to exploit the resulting syntactic annotations to create richer and more extensive linguistic resources. ELDA is organizing two evaluation campaigns of syntactic parsers for French. The first one took place in 2007 while the second one is planned for the end of 2009.

The French project Port-MEDIA will address the multi-domain and multi-lingual robustness and portability of spoken language understanding systems. More specifically, the overall objectives of the project can be summarized as:

- robustness: integration/coupling of the automatic speech recognition component in the spoken language understanding process.
- portability across domains and languages: evaluation of the genericity and adaptability of the approaches implemented in the understanding systems, and development of new techniques inspired by machine translation approaches.
- representation: evaluation of new rich structures for high-level semantic knowledge representation.

ELDA is involved in the production part of the project but also in organizing the evaluation and assessment of speech understanding systems.

## 4.5.2 FBK-irst

[Fondazione Bruno Kessler \(FBK-irst\)](#), formerly Istituto Trentino di Cultura, has been active in the area of evaluation for several years, taking part into the organization of several Evaluation Campaigns. In the following the main evaluation activities carried out in the last years are reported.

- In 2007 FBK-irst and [CELCT](#) launched EVALITA (see section 4.3.3), the first evaluation campaign of NLP tools for Italian. FBK-irst and CELCT are now co-organizing Evalita 2009, the second edition of the campaign.

- FBK is co-chairing, together with the University of Texas at Austin, the ACL SigLex event SemEval-2010.
- FBK-irst has been involved in the organization of RTE (see section 4.4.8) since its inception in 2004. Starting from RTE-2 and up to now, CELCT joined FBK in the organization of the task.
- EUROMATRIXPlus. FBK is a partner of the EuroMatrixPlus project (see 4.3.9).

### 4.5.3 [CELCT](#)

CELCT (Center for the Evaluation of Language and Communication Technologies) is a joint enterprise established by FBK-irst (section 4.5.2 ) and DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz), and funded by the Autonomous Province of Trento. Located in Trento, CELCT began operations in 2003.

The goals of CELCT are to set up infrastructures and develop skills in order to operate successfully in the field of the evaluation of language and communication technologies, becoming a reference point in the field at the national and European levels.

CELCT carries out a variety of activities in the field of HLT evaluation, including the organization of national and international evaluation campaigns, where it is involved in both management tasks and scientific activities. Projects include Evalita, QA@CLEF, TrebleCLEF and RTE. CELCT is also specialized in the creation of linguistic resources, producing and annotating corpora in different languages, from different sources, and at different annotation levels.

### 4.5.4 University of Leeds

The [Centre for Translation Studies of the University of Leeds](#) focuses on the tasks of developing usage scenarios for automated MT evaluation tools. These scenarios will be focussed on possible industrial applications. In particular, in previous work researchers from the University of Leeds showed that N-gram based evaluation scores like BLEU loose sensitivity for higher-quality MT systems, while performance-based metrics, such as metrics based on Information Extraction tasks from MT output, don't show the loss of sensitivity across a wider quality spectrum (Babych and Hartley 2007). These researchers also developed a method for automated error-analysis on the basis of concordance evaluation using BLEU (Babych and Hartley 2008), which is useful for the developers of industrial MT systems. Leeds University's current work concentrates on creating new types of usage scenarios for existing MT evaluation metrics, where the scores can have meaningful interpretation, and possible limitations on the use of automated MT evaluation tools can be avoided.

## 4.6 *Characterization of the Developments of the past 5 years*

There is general consensus that systematic evaluation of HLT technologies has driven research on these technologies. It allows researchers to objectively compare approaches and to reproduce experiments, and more generally to make issues explicit, to validate new ideas and to identify missing science. It is also an important tool to judge funding efficiency and to determine the maturity of the developments for a given application (Geoffrois 2009, p. 61). There is also no doubt that evaluation should remain an important driving force for this research in the coming years. It has even been argued that evaluation can be used to create validated high-

quality linguistic annotations at a relatively low cost (Paroubek 2009, p. 64-65), as a side effect. In short, carrying out systematic evaluation is highly beneficial for the HLT field.

In the past years the focus has been on automating evaluation as much as possible. The methodologies and metrics for automating evaluation are still highly controversial. Many researchers claim that certain metrics that make such automation possible do not provide us with relevant measures for quality assessment. It is to be expected that discussions in this area will continue for some time, and systematic investigation of these matters should be stimulated though it is not to be expected that the issues will be resolved soon.

Many researchers believe that the evaluation efforts are still too fragmented and too scattered. As a remedy, several researchers have argued for a sustainable infrastructure to carry out such evaluations in a systematic manner. For example, Höge (2009, p. 67-68) argues for setting up a Remote Evaluation System Architecture (RESA), inspired by work done in ECESS. Magnini (2009, p. 63) argues for an integrated evaluation framework where single components are not evaluated per se, but rather for their contribution to a global application. In order to achieve this he suggests working on the development of shared communication protocols for single-task components, giving support for interoperability of single-task components within global applications, and setting up a web infrastructure based on web services on the base of these shared communication protocols.

According to Peters (2009, p. 59) evaluation activities cannot operate only via a voluntary networking basis; a solid underlying management and coordination structure is crucial in order to ensure that the programme of activities is viable, consistent and coherent and that it can successfully scale up and embrace new communities and technological paradigms. Höge (2009, p. 68) concretely proposes to set up a “Centre for Remote Evaluation and Development of Language Technologies” which configures and maintains the RESA in accordance with the needs of the community for evaluation and development. Geoffrois, however, warns that the traditional partial grants combined with a lack of dedicated public structures results in a shortage of evaluation infrastructures. Different funding strategies are required to ensure that these infrastructures are suited to the needs of research and development. It is therefore of critical importance to set up new funding strategies for LT evaluation in Europe to get the full benefits of the large investments in the domain. (Geoffrois 2009, p. 62). Special attention, strategies and funding might be required for lesser spoken languages (Vertan 2009).

The model adopted in ECESS might be a good example that implements many basic requirements for such a sustainable infrastructure. The Open Resource Infrastructure (ORI) that will be created in the context of the T4ME Net project, together with the efforts in this project on boosting MT by incorporating and evaluating new types of modules in MT systems, might instantiate such an infrastructure, at least for HLT dealing with multilinguality and machine translation.

A proper set-up of such an infrastructure may have other benefits as well. Such an infrastructure should define and be able to incorporate a set of components and their interfaces for HLT technologies that are agreed upon by the participants in the infrastructure. Access to components provided by others should be as free as possible, perhaps in exchange for contributing ones own (possibly different) components. Sharing components avoids duplication, allows research groups to focus on their speciality, creates *de facto* standards and in fact even actual interoperability! This will allow evaluation to be carried out not only on systems as a whole but also on components of the whole system. Sharing components may require the development of clear agreements, especially to safeguard the interest of commercial parties.

Sharing components with a well-defined and agreed-upon interface may also have the effect that for such components validation can be replaced by the requirement that it should be incorporated in the infrastructure and be able to cooperate with other components in the infrastructure (interoperability), in line with the suggestions made by several validation experts. It will make validation for such components automatic and a by-product of incorporation in the infrastructure.

## 5 Bibliography

Babych, B., Elliott, D., Hartley, A. (2004) Extending MT evaluation tools with translation complexity metrics. In: COLING 2004: the 20th International Conference on Computational Linguistics, p 106-112.

Babych, B., Hartley, A., Elliott, D. (2005) Estimating the predictive power of n-gram MT evaluation metrics across language and text types. In: MT Summit X, p 412-418.

Babych, B., Hartley, A. (2007) Sensitivity of automated models for MT evaluation: proximity-based vs. performance-based methods. In: MT Summit XI Workshop: Automatic procedures in MT evaluation

Babych, B., Hartley, A. (2008) Automated MT evaluation for error analysis: automatic discovery of potential translation errors for multiword expressions. In: ELRA Workshop on Evaluation: Looking into the Future of Evaluation: when automatic metrics meet task-based and performance-based approaches. In conjunction with LREC 2008.

Alan W Black and Keiichi Tokuda. The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. Interspeech 2005.

Braschler, M. and Gonzalo J. (2009), 'Best Practices in System and User Oriented Multilingual Information Access', edited by the TrebleCLEF Consortium, <http://www.trebleclef.eu>

Callison-Burch, C., Osborne, M., Koehn, P. (2006) Re-evaluation the Role of Bleu in Machine Translation Research. In: 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), p 249-256.

Calzolari *et al.* (2009): Proceedings of The *European Language Resources and Technologies Forum: Shaping the Future of the Multilingual Digital Europe*. Vienna, 12th and 13th February 2009 [ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/language-technologies/flarenet-vienna-09-proceedings\\_en.pdf](ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/language-technologies/flarenet-vienna-09-proceedings_en.pdf)

Cieri, Christopher (2006) *What is Quality? Invited Talk at the Workshop on Quality Assurance and Quality Measurement for Language and Speech Resources*. In LREC 2006: Fifth International Conference on Language Resources and Evaluation, [http://papers.ldc.upenn.edu/LREC2006/Quality\\_Workshop.ppt](http://papers.ldc.upenn.edu/LREC2006/Quality_Workshop.ppt)

Dickinson, M., C. Min Lee (2008) Detecting Errors in Semantic Annotations. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco, 28-30 May 2008, p. 605-610

Doddington, G. (2002) "Automatic evaluation of machine translation quality using n-gram cooccurrence statistics". *Proceedings of the Human Language Technology Conference (HLT)*, San Diego, CA pp. 128—132

Galliano, S. Gravier, G., Chabard L. (2009) The ESTER 2 Evaluation Campaign for Rich Transcription of French radio Broadcasts. In *INTERSPEECH 2009*, p. 2583-2586

Geoffrois, E. (2009), "Language Technology Evaluation: which Funding Strategy?", in Calzolari *et al* (2009), pp. 61-62 and [http://www.flarenet.eu/sites/default/files/Geoffrois\\_Presentation.pdf](http://www.flarenet.eu/sites/default/files/Geoffrois_Presentation.pdf)

Van den Heuvel, Henk (2009) The "Standard Deviation" of LR Quality, in Calzolari *et al* (2009), pp. 53-54 and: [http://www.flarenet.eu/sites/default/files/van\\_den\\_Heuvel\\_Presentation.pdf](http://www.flarenet.eu/sites/default/files/van_den_Heuvel_Presentation.pdf)

Höge, H. (2009) "Proposal to launch a Support Centre for Remote Evaluation and Development of Language Technologies", in Calzolari *et al.* (2009), p. 67-68 and [http://www.flarenet.eu/sites/default/files/Hoege\\_Presentation.pdf](http://www.flarenet.eu/sites/default/files/Hoege_Presentation.pdf).

de Kok, D., Ma, J. and van Noord G. (2009) 'A generalized method for iterative error mining in parsing results'. In: *ACL2009 Workshop Grammar Engineering Across Frameworks (GEAF)*, Singapore, 2009. See <http://www.let.rug.nl/vannoord/papers/geaf2009.pdf>

Lavie, A., Sagae, K. and Jayaraman, S. (2004) "The Significance of Recall in Automatic Metrics for MT Evaluation" in *Proceedings of AMTA 2004, Washington DC. September 2004*

van Leeuwen, D., Kessen, J. (2008) *Evaluation plan for the North- and South-Dutch Benchmark Evaluation of Speech recognition Technology*. Available at <http://en.scientificcommons.org/4225707> (last access 2009/10/22)

Lin, Chin-Yew. 2004. *ROUGE: a Package for Automatic Evaluation of Summaries*. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.

Magnini, B. (2009), "Toward an Integrated Evaluation Framework", in Calzolari *et al.* (2009), p. 63 and

[http://www.flarenet.eu/sites/default/files/Magnini\\_Presentation.pdf](http://www.flarenet.eu/sites/default/files/Magnini_Presentation.pdf)

Moreau, N. (2009) 'Best Practices in Language Resources for Multilingual Information Access, edited by the TrebleCLEF Consortium, <http://www.trebleclef.eu>

Navigli, R. (2006) Consistent Validation of Manual and Automatic Sense Annotations with the Aid of Semantic Graphs. In *Computational Linguistics*, vol. 32 (2), p. 273-281

van Noord, G. (2004), 'Error mining for Wide-Coverage Grammar Engineering', *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004, Barcelona, Spain. See [http://acl.ldc.upenn.edu/acl2004/main/pdf/82\\_pdf\\_2-col.pdf](http://acl.ldc.upenn.edu/acl2004/main/pdf/82_pdf_2-col.pdf)

Papineni, K. *et al.* (2002) *BLEU: a Method for Automatic Evaluation of Machine Translation*, Proceedings of ACL 2002, 311-318. <http://acl.ldc.upenn.edu/acl2002/MAIN/pdfs/Main076.pdf>

Paroubek, P. (2009) "Evaluation: A paradigm that produces high quality language resources", in Calzolari *et al.* (2009), p. 64-66.

Peters, C. (2008) *What happened in CLEF 2008. Introduction to the Working Notes*. Available at [http://www.clef-campaign.org/2008/working\\_notes/CLEF2008WN-Contents.html](http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html) (last access 2009/10/22)

Peters, C. (2009a) *Evaluation of Technology for Multilingual Information Access: the Next Step*, in Calzolari *et al.* (2009), 57-59 and [http://www.flarenet.eu/sites/default/files/Peters\\_Presentation.pdf](http://www.flarenet.eu/sites/default/files/Peters_Presentation.pdf)

Peters, C. (2009b) *What happened in CLEF 2009. Introduction to the Working Notes*. Available at [http://www.clef-campaign.org/2008/working\\_notes/CLEF2009WN-Contents.html](http://www.clef-campaign.org/2008/working_notes/CLEF2009WN-Contents.html) (last access 2009/10/22).

C. J. van Rijsbergen (1979). *Information retrieval*. Butterworths, London, 2nd edition.

Sanderson M. and Braschler M. (2009) 'Best Practices for Test Collection Creation and Information Retrieval System Evaluation, edited by the TrebleCLEF Consortium, <http://www.trebleclef.eu>

Sagot, B. and de la Clergerie, E. (2006) 'Error mining in parsing results'. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 329–336, Morristown, NJ, USA.

Schiel, F. (2009) Towards more effective LR validation, in Calzolari *et al.* (2009), pp 55-56 and [http://www.flarenet.eu/sites/default/files/Schiel\\_Presentation.pdf](http://www.flarenet.eu/sites/default/files/Schiel_Presentation.pdf)

Snover, M. *et al.* (2006) *A Study of Translation Edit Rate with Targeted Human Annotation*. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-2006), pages 223–231, Cambridge, MA, August.

Spoutsova, D., Pecina, P., Hajic J., Spoutsa, M. (2008) Validating the Quality of Full Morphological Annotation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco, 28-30 May 2008, p. 1132-1135

Thurmair, G. (2007) Automatic evaluation in MT system production. In: MT Summit XI Workshop: Automatic procedures in MT evaluation, 11 September 2007, Copenhagen, Denmark.

Vertan, C. (2009), "Evaluation of HLT-tools for less-spoken languages", in Calzolari *et al.* (2009), pp. 69-70 and [http://www.flarenet.eu/sites/default/files/Vertan\\_Presentation.pdf](http://www.flarenet.eu/sites/default/files/Vertan_Presentation.pdf)

## **A Evaluation Packages available at ELRA**

The following packages that may be used for HLT evaluations can be obtained via ELRA:

- AURORA 2 Project Database
- AURORA 3 Project database
- AURORA 4 Project Database
- E0002 TC-STAR 2005 Evaluation Package - ASR English
- E0003 TC-STAR 2005 Evaluation Package - ASR Spanish
- E0004 TC-STAR 2005 Evaluation Package - ASR Mandarin Chinese
- E0005 TC-STAR 2005 Evaluation Package - SLT English-to-Spanish
- E0006 TC-STAR 2005 Evaluation Package - SLT Spanish-to-English
- E0007 TC-STAR 2005 Evaluation Package - SLT Chinese-to-English
- E0008 The CLEF Test Suite for the CLEF 2000-2003 Campaigns
- E0009 CHIL 2004 Evaluation Package
- E0010 CHIL 2005 Evaluation Package
- E0011 TC-STAR 2006 Evaluation Package - ASR English
- E0012-01 TC-STAR 2006 Evaluation Package - ASR Spanish – CORTES
- E0012-02 TC-STAR 2006 Evaluation Package - ASR Spanish – EPPS
- E0013 TC-STAR 2006 Evaluation Package - ASR Mandarin Chinese
- E0014 TC-STAR 2006 Evaluation Package - SLT English-to-Spanish
- E0015-01 TC-STAR 2006 Evaluation Package - SLT Spanish-to-English – CORTES
- E0015-02 TC-STAR 2006 Evaluation Package - SLT Spanish-to-English – EPPS
- E0016 TC-STAR 2006 Evaluation Package - SLT Chinese-to-English
- E0017 CHIL 2006 Evaluation Package
- E0018 ARCADE II Evaluation Package
- E0019 CESART Evaluation Package
- E0020 CESTA Evaluation Package
- E0021 ESTER Evaluation Package
- E0022 EQueR Evaluation Package
- E0023 EvaSy Evaluation Package
- E0024 MEDIA Evaluation Package

- E0025 TC-STAR 2007 Evaluation Package - ASR English
- E0026-01 TC-STAR 2007 Evaluation Package - ASR Spanish – CORTES
- E0026-02 TC-STAR 2007 Evaluation Package - ASR Spanish – EPPS
- E0027 TC-STAR 2007 Evaluation Package - ASR Mandarin Chinese
- E0028 TC-STAR 2007 Evaluation Package - SLT English-to-Spanish
- E0029-01 TC-STAR 2007 Evaluation Package - SLT Spanish-to-English – CORTES
- E0029-02 TC-STAR 2007 Evaluation Package - SLT Spanish-to-English – EPPS
- E0030 TC-STAR 2007 Evaluation Package - SLT Chinese-to-English
- E0031 TC-STAR 2006 Evaluation Package End-to-End
- E0032 TC-STAR 2007 Evaluation Package End-to-End
- E0033 CHIL 2007 Evaluation Package
- E0034 EASy Evaluation Package
- W0013 TSNLP (Test Suites for NLP Testing)
- W0029 Amaryllis Corpus - Evaluation Package

## B List of Acronyms and Project Names

Acronym	(English) Expansion	URL
ACE	Automatic Content Extraction French-speaking Speech Communication	<a href="http://www.itl.nist.gov/iad/mig/tests/ace/ace07/">http://www.itl.nist.gov/iad/mig/tests/ace/ace07/</a>
AFCP	Association Italian Association for Artificial Intelligence	<a href="http://www.isca-students.org/afcp_association_francophone_de_la_communication_parle">http://www.isca-students.org/afcp_association_francophone_de_la_communication_parle</a> <a href="http://www.afcp.org/">http://www.afcp.org/</a>
AI*IA		<a href="http://www.aixia.it/">http://www.aixia.it/</a>
ALBAYZÍN		<a href="http://www.rthabla.es/">http://www.rthabla.es/</a>
ARCADEII	Evaluation of bilingual text and vocabulary alignment systems. Advanced Research and Development	<a href="http://elda.org/article135.html">http://elda.org/article135.html</a>
ARDA	Activity	<a href="http://www.ic-arda.org/">http://www.ic-arda.org/</a>
ASR	Automatic Speech Recognition	
BAS	Bavarian Archive for Speech Signals BAS Services	<a href="http://www.phonetik.uni-muenchen.de/Bas/BasValideng.html">http://www.phonetik.uni-muenchen.de/Bas/BasValideng.html</a> <a href="http://www.bas-services.de/">http://www.bas-services.de/</a>
BASSS	Schiel	
BITS	<u>B</u> AS Infrastructures for <u>T</u> echnical <u>S</u> peech Processing	<a href="http://www.bas.uni-muenchen.de/Forschung/BITS/">http://www.bas.uni-muenchen.de/Forschung/BITS/</a>
BLEU	Bilingual Evaluation Understudy	
Blizzard Challenge		<a href="http://festvox.org/blizzard/">http://festvox.org/blizzard/</a>
CEA-LIST	CEA System and Technology Integration Laboratory Atomic Energy	<a href="http://www-list.cea.fr/index.htm">http://www-list.cea.fr/index.htm</a>
CEA-LIST	Commissariat Center for the Evaluation of Language and Communication	<a href="http://www.cea.fr/">http://www.cea.fr/</a>
CELCT	Technologies Evaluation of terminology extraction tools, including tools for extracting ontologies and semantic relations	<a href="http://www.celct.it/">http://www.celct.it/</a>
CESART	Evaluation of Machine	<a href="http://elda.org/article137.html">http://elda.org/article137.html</a>
CESTA	Translation Systems Computers in the Human Interaction	<a href="http://elda.org/article136.html">http://elda.org/article136.html</a>
CHIL	Loop	<a href="http://chil.server.de/">http://chil.server.de/</a>
CLARIN	Common Language Resources and Technology	<a href="http://www.clarin.eu/">http://www.clarin.eu/</a>

Infrastructure

CLEF	Cross Lingual Evaluation Forum	<a href="http://www.clef-campaign.org/">http://www.clef-campaign.org/</a>
CMU	Carnegie Mellon University	<a href="http://www.cmu.edu/index.shtml">http://www.cmu.edu/index.shtml</a>
CST	Center for Sprogteknologi, Copenhagen, Denmark	<a href="http://cst.dk/validation/index.html">http://cst.dk/validation/index.html</a>
DARPA	Defense Advanced Research Projects Agency	<a href="http://www.darpa.mil/">http://www.darpa.mil/</a>
DFKI	Deutsches Forschungszentrum für Künstliche Intelligenz	<a href="http://www.dfki.de/">http://www.dfki.de/</a>
DGA/CEP	French Defense expertise and test center for speech and language processing	<a href="http://www.defense.gouv.fr/dga">http://www.defense.gouv.fr/dga</a>
DUC	Document Understanding Conference	<a href="http://duc.nist.gov/">http://duc.nist.gov/</a>
EAGLES	Expert Advisory Group for Language Engineering Standards	<a href="http://www.ilc.cnr.it/EAGLES96/home.html">http://www.ilc.cnr.it/EAGLES96/home.html</a>
EASY	An evaluation campaign designed to test syntactic parsers.	<a href="http://elda.org/article138.html">http://elda.org/article138.html</a>
ECESS	European Center of Excellence for Speech Synthesis	<a href="http://www.ecess.eu/">http://www.ecess.eu/</a>
ELDA	Evaluations and Language Distribution Agency	<a href="http://www.elda.org/">http://www.elda.org/</a>
ELRA	European Language Resource Association	<a href="http://www.elra.info/">http://www.elra.info/</a>
EQUER	Evaluation of Question/Answerin g systems.	<a href="http://elda.org/article139.html">http://elda.org/article139.html</a>
ESTER	Evaluation of automatic broadcast news transcriptions systems	<a href="http://elda.org/article140.html">http://elda.org/article140.html</a>
EU	European Union	
EUROMATRIX		<a href="http://www.euromatrix.net/">http://www.euromatrix.net/</a>
EUROMATRIXPlus		<a href="http://www.euromatrixplus.net/">http://www.euromatrixplus.net/</a>
EvalDA		<a href="http://www.evalda.org/rubrique25.html">http://www.evalda.org/rubrique25.html</a>
Evalita	Evaluation of NLP and Speech Tools for Italian	<a href="http://evalita.fbk.eu/">http://evalita.fbk.eu/</a>
EVASY	Evaluation of Speech synthesis systems	<a href="http://elda.org/article141.html">http://elda.org/article141.html</a>
FBK-irst	Fondazione Bruno Kessler	<a href="http://www.fbk.eu/">http://www.fbk.eu/</a>

	Framework for Machine Translation Evaluation within the ISLE initiative	<a href="http://www.isi.edu/natural-language/mteval/">http://www.isi.edu/natural-language/mteval/</a>
FEMTI		<a href="http://festvox.org/index.html">http://festvox.org/index.html</a>
Festvox		<a href="http://www.flarente.eu/">http://www.flarente.eu/</a>
FLaReNet	Fostering Language Resources Network	
FP6	6th Framework	
FP7	7th Framework	
	Global Autonomous Language Exploitation	<a href="http://www.darpa.mil/ipto/programs/gale/gale.asp">http://www.darpa.mil/ipto/programs/gale/gale.asp</a>
GALE	Gengo-Shigen- Kyokai ( “Language Resources Association”), Japan	<a href="http://www.gsk.or.jp/index_e.html">http://www.gsk.or.jp/index_e.html</a>
GSK		
HLT	Human Language Technology	
iCLEF	Interactive Cross- Language Retrieval	<a href="http://nlp.uned.es/iCLEF/">http://nlp.uned.es/iCLEF/</a>
ICT	Information and Communication Technology INformation, Filtering, Evaluation	<a href="http://www.infile.org/">http://www.infile.org/</a>
INFILE		<a href="http://www.ilc.cnr.it/EAGLES/isle/right.html">http://www.ilc.cnr.it/EAGLES/isle/right.html</a>
ISLE	International Standards for Language Engineering	
ISST	Italian Syntactic- Semantic Treebank), International Workshop on Spoken Language Translation	<a href="http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=894/vers=ita">http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=894/vers=ita</a>
IWSLT		<a href="http://iwslt07.fbk.eu/">http://iwslt07.fbk.eu/</a>
LDC	Linguistic Data Consortium	<a href="http://www ldc.upenn.edu/">http://www ldc.upenn.edu/</a>
LE-System	Language Engineering System	
LR	Language Resource	
LREC	Language Resource and Evaluation Conference	<a href="http://www.lrec-conf.org/">http://www.lrec-conf.org/</a>
	Large Vocabulary Continuous Speech Recognition	
LVCSR	Speech and Language Technologies for Arabic	<a href="http://www.nemlar.org/">http://www.nemlar.org/</a>
MEDAR	Evaluation of Man- Machine dialogue systems	<a href="http://elda.org/article115.html">http://elda.org/article115.html</a>
MEDIA	The METEOR Automatic Machine Translation	
METEOR	Evaluation System	<a href="http://www.cs.cmu.edu/afs/cs/user/alavie/WWW/METEOR/">http://www.cs.cmu.edu/afs/cs/user/alavie/WWW/METEOR/</a>
ML	Machine Learning	

MOS	Mean Opinion Score	
MT	Machine Translation	
Multext	Multilingual Text Tools and Corpora	<a href="http://aune.lpl.univ-aix.fr/projects/multext/">http://aune.lpl.univ-aix.fr/projects/multext/</a>
N-BEST	Northern and Southern Dutch Benchmark Evaluation of Speech recognition Technology	<a href="http://speech.tn.tno.nl/n-best/">http://speech.tn.tno.nl/n-best/</a>
NE	Named Entity Detection	
NII-SRC	Speech Resources Consortium, Japan	<a href="http://research.nii.ac.jp/src/eng/index.html">http://research.nii.ac.jp/src/eng/index.html</a>
NIST	National Institute of Standards and Technology, US	<a href="http://www.nist.gov/index.html">http://www.nist.gov/index.html</a>
NIST metric	BLEU-derived MT evaluation metrics designed by NIST researchers	
NLP	Natural Language Processing	
NTCIR	NII Test Collection for IR Systems	<a href="http://ntcir.nii.ac.jp/">http://ntcir.nii.ac.jp/</a>
Oriental	Multilingual access to interactive communication services for the Mediterranean and the Middle East	<a href="http://www.speechdat.org/ORIENTEL/index.html">http://www.speechdat.org/ORIENTEL/index.html</a>
Oriental	Open Resource Infrastructure	
PAROLE	PAROLE project Pattern Analysis, Statistical Modelling and Computational Learning	<a href="http://www.ub.edu/gilcub/SIMPLE/simple.html">http://www.ub.edu/gilcub/SIMPLE/simple.html</a>
PASCAL	PASCAL Successor Project	<a href="http://www.pascal-network.org/">http://www.pascal-network.org/</a>
PASCAL2	Evaluation of	<a href="http://www.pascal-network.org/">http://www.pascal-network.org/</a>
PASSAGE-2	Dependency Parsing	<a href="http://atoll.inria.fr/passage/index.en.html">http://atoll.inria.fr/passage/index.en.html</a>
PER	Position-independent word error rate	
PORT-MEDIA		<a href="http://port-media.org/doku.php">http://port-media.org/doku.php</a>
QQC	Quick Quality Check	
R&D	Research & Development	
RESA	Remote Evaluation System Architecture	
ROUGE	Recall-Oriented Understudy for Gisting Evaluation	<a href="http://en.wikipedia.org/wiki/ROUGE_(metric)">http://en.wikipedia.org/wiki/ROUGE_(metric)</a>
RTE	Recognizing Textual Entailment	<a href="http://pascallin.ecs.soton.ac.uk/Challenges/RTE3/">http://pascallin.ecs.soton.ac.uk/Challenges/RTE3/</a>
SED	Sound Event Tracking	

Semeval	Semantic Evaluation	<a href="http://semeval2.fbk.eu/semeval2.php">http://semeval2.fbk.eu/semeval2.php</a>
Senseval	Sense Disambiguation	<a href="http://www.senseval.org/">http://www.senseval.org/</a>
SIMPLE	Evaluation	<a href="http://www.ub.edu/gilcub/SIMPLE/simple.html">http://www.ub.edu/gilcub/SIMPLE/simple.html</a>
SITEC	SIMPLE Project Speech Information TEchnology & Industry Promotion Center, Wonkwang University, Korea	<a href="http://www.sitec.or.kr/English/">http://www.sitec.or.kr/English/</a>
SLR	Spoken Language Resources	
SMT	Statistical Machine Translation	
SNR	Signal to Noise Ratio	
SpeechDat(E)	Eastern European Speech Databases for Creation of Voice Driven Teleservices	<a href="http://www.fee.vutbr.cz/SPEECHDAT-E/">http://www.fee.vutbr.cz/SPEECHDAT-E/</a>
SpeechDat(II)	Databases for the Creation of Voice Driven Teleservices	<a href="http://www.speechdat.org/SpeechDat.html">http://www.speechdat.org/SpeechDat.html</a>
SpeechDat-Car	Speech Databases recorded in Vehicles	<a href="http://www.speechdat.org/SP-CAR/">http://www.speechdat.org/SP-CAR/</a>
SpeeCon	Speech-driven Interfaces for Consumer Devices	<a href="http://www.speechdat.org/speecon/index.html">http://www.speechdat.org/speecon/index.html</a>
SPEX	SPeech EXpertise centre, the Netherlands	<a href="http://www.spex.nl/validationcentre/">http://www.spex.nl/validationcentre/</a>
SRL	Speaker Diarization	
SST	Speech-to-Speech Translation	
STEVIN	Essential Speech and Language Technology Resources for the Dutch Language	<a href="http://taaluniversum.org/taal/technologie/stevin/">http://taaluniversum.org/taal/technologie/stevin/</a>
STT	Speech-to-Text	
SVL	Speaker Tracking Technologies for Multilingual Europe	
T4ME Net	Network Text Analysis	
TAC	Conference	<a href="http://www.nist.gov/tac/">http://www.nist.gov/tac/</a>
TC-STAR	TC-STAR Project	<a href="http://www.elda.org/en/proj/tcstar-wp4/">http://www.elda.org/en/proj/tcstar-wp4/</a>
TER	Translation Error Rate	
TIDES	Translingual Information Detection Extraction and Summarization	
TIPSTER	DARPA TIPSTER Project	
TrebleCLEF	TrebleCLEF Project	<a href="http://www.trebleclef.eu/">http://www.trebleclef.eu/</a>
TREC	Text REtrieval Conference	<a href="http://trec.nist.gov/">http://trec.nist.gov/</a>

TRS	Speech	
TTS	Transcription	
TUT	Text-to-Speech	
	Turin University	<a href="http://www.di.unito.it/~tutreeb/">http://www.di.unito.it/~tutreeb/</a>
	Treebank	
US	United States of	
	America	
VC	Validation Centre	
VCom	(ELRA) Validation	<a href="http://www.elra.info/Validation.html">http://www.elra.info/Validation.html</a>
	Committee	
VP	Verb Phrase	
W3C	World Wide Web	
	Consortium	<a href="http://www.w3.org/">http://www.w3.org/</a>
WER	Word Error Rate	
WLR	Written Language	
	Resource	
XML	eXtensible Mark-up	
	Language	<a href="http://www.w3.org/XML/">http://www.w3.org/XML/</a>