Project ref. no.:	LE4-8340
Project title:	Evaluation in Language an Speech Engineering.
Deliverable status:	Public
Contractual date of delivery:	April 30th 1999
Actual date of delivery:	
Deliverable number:	D1.2
Deliverable title:	"Follow-up Evaluation Proposal"
Type:	Report
Status:	Pre-Final version 1.1
Number of pages:	18
WP contributing to the deliverable:	WP1
WP / Task responsible:	Limsi - CNRS, Bâtiment 508 Université Paris XI Dépt. Communication Homme Machine, BP 133 - 91403 ORSAY Cedex
Author(s):	Editor: Patrick Paroubek. Contributors: Niels Ole Bernsen, Marc Blasband, Nicoletta Calzolari, Jean-Pierre Chanod, Khalid Choukri, Laila Dybkjær, Robert Gaizauskas, Steven Krauwer, Isabelle de Lamberterie, Joseph Mariani, Klaus Netter, Patrick Paroubek, Martin Rajman, Antonio Zampolli
EC Project Officer:	Giovanni Battista Varile
Keywords:	EVALUATION, QUANTITATIVE, BLACK BOX, NATURAL LANGUAGE PROCESSING, MULTILINGUALITY, CONTROL TASK, PROPOSAL, FP5.

Abstract: In the following, we will retrace the efforts made by the ELSE consortium to propose follow-up evaluation activities to the ELSE project. We will start with the initial list of 31 control task identified at the beginning of the project as deemed of worth of attention given the current level of development achieved by Language Technologies. Then we will consider the integration of evaluation in the call for proposal scheme used by the Commission. After which we will explain the reasons that made us favor as best candidates to start an evaluation program six control tasks out of the previous 31, and how these control tasks could be merged into a single one: NODE (News On Demand Evaluation).Finally, we will present the EvaLE proposal of which ELSE was a contributor and which has been submitted to the first call of FP5.

Contents

3,4	S&T	Machine Translation	[DARPA/USA/92/93/94]
5,6	S&T	Multilingual data alignment	[ARC A2/Aupelf/FR/95]
7	T	Terminology Extraction	[ARC A3 /Aupelf/FR/95]
8,9	S&T	Document Extraction	[TREC/DARPA/USA/92-98]
10,11	S&T	Language Understanding	[MUC/DARPA/USA/87-97]
12	T	Text Generation (from information templates)	.
13	T	Summary Generation	[DARPA/USA/98]
14	T	Text Segmenting	.
15	S	Speech Segmenting	.
16	S	Speech Recognition	[DARPA/USA/84-98] and [ARC B1/Aupelf/FR/95]
17	S	Speech Synthesis	[ARC B3Aupelf/FR/95]
18,19	S&T	Topic detection & Tracking	[DARPA/USA/98]
20	T	Part-Of-Speech Tagging	[GRACE-CNRS/FR/94-98]
21	T	Parsing	[Parseval/USA/92] and [SPARKLE/EU/96]
22	T	Lemmatizers	[Morpholympics/Germany/94]
23	T	Word Sense Disambiguation	[SENSEVAL/98]
24	T	Predicate Argument Structure	.
25,26	S&T	Coreference Identification	[DARPA/USA/95+98]
27,28	S&T	Named Entity Extraction	[DARPA/USA/95+98]
29	S	Database Dialogue Querying	[EuroSpeech97/ELSNET/97] and [ARC B2/Aupelf/FR/95]
30	T	Hand Written Recognition	[NIST/USA/92]
33	S	Speaker Verification / Recognition	[NIST/USA/96/97/98]
31	S	Language Identification	

ELSE uses as reference frame, the generic abstract architecture of a cross-language intelligent information extraction system, which can access both local and distributed databases. In this context information extraction is meant in a broad sense, encompassing both the classical meanings of Information Extraction (IE), i.e. template filling from documents, and Information Retrieval (IR), i.e. document selection. Such system would have multi-modal input and output and would be able to intelligently adapt its behavior to a particular query, for instance by choosing between classical IE and IR functionality, or deciding whether to consult either a local database or the WEB to access information sources from various media. Our abstract application can be seen as a sort of super information browser/finder.

This generic architecture was helpful for determining the list of control tasks that we propose here. Each evaluation task corresponds to an abstract functionality or module of the architecture. The various components of the architecture can be developed along the following 3 dimensions:

1. Information Profiling (data analysis).
2. Information Querying (dialog management issues and mapping results to query).
3. Information Presentation (output modality selection, language generation).

Evaluation points can be selected at the input and output of individual modules of such architecture and also at any point along arbitrary module chains. Thus, new evaluation tasks can be defined by linking various modules of the abstract architecture in a braided fashion [KN95]. Our abstract architecture is very

spoken in Europe would be totally impractical.

Thus we must make either a drastic and arbitrary selection (by necessity not entirely based on scientif criteria) or find a way to generalize the results obtain for a given technology in one language to other languages. An alternative solution to reduce the number of languages addressed while retaining roughly the same language coverage for an evaluation campaign would be to set cross language functionality requirements for the control task (specifying different input and output languages, e.g. in Information Retrieval). But this solution does not apply to taks which are intrinsically monolingual like speech generation.

The solution that was finally taken in the follow-up proposal (see section ???) was to require that each participant addresses at least two languages (their own and another European language), and that for any evaluation campaign there is at least one language common to all participants, and at least two participants for any language. Such scheme was implemented in the SQALE [SYLL97] project. It enables some generalization of the evaluation results in the case where a system A obtains the best results for his language and has better results than a system B on the pivotal language, it is expected that the system A will have better results than B when addressing any other language in the language lineage (e.g. French, Spanish, Italian, Portuguese and Romanian all derive from Latin) of either the pivotal language or the language specific to system A.

An extra factor of homogeneity, would for all the evaluation campaigns to share a common language. (American) English is a strong candidate, since it is spoken and understood by a large number of people, it represents a large market and given possible co-operation activities between the EU and the US in the field of LE evaluation.

4. Proactive or Reactive Approach?

Depending on whether a proactive or a reactive solution is sought, the difference in strategy reflects the disparity of requirements imposed by each type of solution. With the former option, a list of topics is defined in advance of their publication in a unique call for proposals (asking for both evaluators and participants). With the latter option, the evaluation topics are determined by the contents of the selected projects from a first call and a subsequent call is needed to select the evaluators.

If the proactive solution is chosen, the call for proposals should ask either for consortia for each of the evaluation topics, or for larger consortia covering the full set of topics. The first solution is lighter to implement, but the second one allows for a better overall infrastructure more apt to co-ordinate the various evaluations of components and complete systems, but is harder to manage (70 participants or more). The consortia should include a set of organizers for managing the evaluation campaigns in one (their own) or several languages. Each proposal should consider the common language and up to 3 other languages. It should include the description of the way the consortium plans to organize the campaign, the Language Resources (LR) that will be used for training and testing the systems, their cost, and their providers (who will participate as subcontractors), the list of potential participants for each language (at least two), who will also participate as subcontractors. We strongly suggest that the permanent European evaluation organization mentioned before should be a partner in the final consortia in order to capitalize on the results of the different evaluation campaigns. Having the LR providers and the participants as subcontractors allows for more flexibility (in case of reduction in the number of participants down to two or if a change in the participant list occurs). Alternatively, if the cost is too high to support all the potential participants, the consortia could first select a set of participants based on the evaluation results obtained in a dry run. Second, the consortia would finance only the best systems for the final test, up to a certain number. Each participant would receive a fixed amount of resources corresponding to the estimated cost of the participation in the evaluation campaign (typically for adapting his system to the test conditions).

If the reactive solution is chosen, the evaluation topics are determined by the content of the selected projects, which perforce address evaluation as a complementary issue. Subsequently the selected projects

of the features of the control task should be done in order to allow the largest number of systems to participate.

Although a parallel of some sort could be drawn, evaluation activities should not be mistakenly put on a same standing as concertation and dissemination activities. In particular, organizing an evaluation requires the ability to maintain high bandwidth communication with the participants on highly technical grounds, e.g. in order to finalize the evaluation metrics. While concertation activities can be successfully achieved with much lighter means judiciously distributed over time.

A mixed solution between the purely proactive and purely reactive solutions is possible. Some evaluation topics could be selected beforehand and published in the first call for proposal, while others could be defined according to the projects selected after the first call. The ELSE consortium favors the proactive approach and a single consortium, constituted as a network or evaluation organizers, with the support of the permanent European evaluation organization.

Note that if the classical EU contract scheme is used to implement evaluation, and if the participants are funded, with the evaluator as only the evaluator (all the participants and the corpus providers are his subcontractors), then the usual limitation imposed on the amount of resources devoted to "third party assistance" should be modified or waived. The amount of resources could exceed the allowed value, just because of the number of participants or the cost of the linguistic data needed for evaluation.

5. Integrating Evaluation in the Call for Proposals.

In order to include evaluation in the FP5 agenda, it is proposed to include this topic in the first call for proposals. Evaluation campaigns would have a 2-year duration, in order to allow for more progress and research work between two campaigns than in the DARPA ones. We expect a certain amount of delay in deploying the paradigm of evaluation because it will be the first time that it will be used on such scale and in the context of EU programs (3 years were necessary for DARPA to go from the drawing board to a real implementation for the speech recognition campaigns).

In a proactive scheme, the topics (related to both written and spoken language processing) should be selected beforehand and included in the call for proposals. These topics should cover both complete systems and systems components, and should have links between them, thus allowing the progress obtained in one field to influence the development of another field. A straightforward way of implementing these links between topics is to have part of the evaluation data that is common to related topics, and therefore in the same language. They should be of interest for LE research, but also for LE industry. To that extent they should be proposed by a scientific committee and submitted to the appreciation of an industrial panel.

As a fallback option, there exists the possibility of including an evaluation task in each candidate project. The evaluation task would constitute a sort of concertation activity where provision would be made for the needs of an evaluation campaign. The resources needed could be contracted out or produced by a subset of the concerned projects. The possible evaluation topics would be determined by the nature of research and technology projects running at a given time, maybe according to project (possibly technology based) clusters, different from the existing project clusters, which are inspired by market considerations. In that case, management becomes more difficult because it is more distributed. It still requires a coordinating entity, which could be as a last resort a specific project. In this reactive scheme driven by the content of the accepted proposals, we may lose the benefits of capitalizing on the evaluation expertise over a long period of time, as there is no guaranty of continuity of the project content across framework programs.

6. Six Candidate Control Tasks for Technology Evaluation.

The main areas of language engineering that are current central preoccupations of researchers and

Language Model Evaluation		X
Technique	X	X

	Text	Speech	Image	Mono/Multilingual
Broadcast News		X		Mono
Cross-Lingual Information Retrieval / Extraction	X			Multi
Text To Speech Synthesis		X		Mono
Text Summarization	X			Mono
Language Model Evaluation	X	X		Mono
Technique	X			Mono

This table shows the relation between the six control tasks and their multimedia and multilinguality aspects.

Naturally, the previous list contains very broadly scoped control tasks. According to the needs, the tasks could be refined into more specific subtasks, or implemented in conjunction with other correlated subsidiary control tasks.

The following table presents the different types of data needed to implement the six control tasks under consideration.

Control Task	*By-product Data Resources*
Broadcast News	Text transcription of speech signal (possibly time-aligned).
Cross-Lingual Information Retrieval / Extraction	Multilingual query/document pairs.
Text To Speech Synthesis	Speech signal for a text
Text Summarization	Document and summary pairs.
Language Model Evaluation	Word predictions (e.g. probability tagging).
Technique	Text with Part-Of-Speech tagging, Lemmatization, Syntactic annotation and Word Sense tagging.

The evaluation of these six tasks will produce data resources (see table above). Out of the 36 reuse possibility of these resources between the six control tasks, 22 are actually possible. Each time, the data

Evaluation criteria will have to found based on the sole of embedded module functionality (e.g. processing speed, language coverage etc.).

The differences between language, culture, environment and application will be parameters of the comparison process.

1. Transcription of Minutes of Virtual Meetings

We propose to use the following usage criteria as comparison points:

- The readability of the transcription;
- The navigation through the written minutes;
- The identification of the speakers;
- The restrictions that the systems put on the meeting itself;
- Comparing the quality with human made minutes;
- The general usefulness of the transcription.

Ideally the test will be the transcriptions of real useful meetings with participants who want to achieve something during that meeting.

2. Multimedia Tourist Information

Multimedia tourist information systems can be compared with the following user-oriented criteria:

- The readability of the output in the user's language;
- The adequacy of the images and sound to the message;
- The precision of the search;
- The completeness of the search;
- The navigation between solutions;
- The number of trips, reservation, purchases made through the application;

3. Text to Speech for the Blind

Text to speech systems for the blind can be compared with the following user-oriented criteria:

- The understandability of the dictation;
- The navigation through the text in relative (e.g. 'repeat the previous paragraph') and exact (e.g. 'read chapter 4') terms;
- The general pleasantness of the dictation (measured e.g. by how long does the user listen without interruption);
- Comparing the quality with human made text to speech;
- The usefulness of the dictation for the users.

4. Summarization

Deployed summarization systems for a given domain or application can be compared with the following user-oriented criteria:

- Readability and understandability;
- Perceived correctness (ambiguities are resolved by the user when he reads);
- Style of the output;
- Usefulness;
- The complexity of language used in the input documents.

8. One Control Task: NODE (News On Demand Evaluation).

1. An evaluation protocol specification, including control task specification (the task to be performed by the systems being evaluated), metrics, data representation formalisms, and the relevant documentation.
2. Development data representative of the control task and in sufficient amount to enable a full validation of the evaluation protocol

The control task which has been selected as presenting interesting features both for Speech and Text analysis is "News on Demand" i.e. indexing, search and interactive browsing of news material, using raw broadcast data (either radio or TV). Using a very coarse grained description; this task includes sub-tasks such as:

1. Sound source separation
2. Speaker identification
3. Speech transcription in noisy environments with multiple speakers,
4. Named entity recognition,
5. Topic detection and tracking
6. Information extraction
7. Spoken document indexing and retrieval.

To address Multilinguality, the project plans to handle a common pivotal language (American English) on which all partners will test their technologies besides 3 other languages (French, Italian and Dutch).

EPs will be validated by running full-fledged evaluation campaigns among the project participants, who will provide baseline reference results, while testing the technologies which are part of their recognized domain of expertise.

To initiate the deployment of evaluation on a larger scale at the European level, EvaLE plan to open its second phase of evaluation campaign to a limited number of participant from outside the project, on a basis of 1 per consortium member. This will complement dissemination activities in making the paradigm of evaluation known outside of the project and hopefully initiate a wider deployment of evaluation that could be supported by FP5 clustering activities. The objective of EvaLE is not only to produce Evaluation Packages, but also to initiate the spreading of the paradigm of evaluation in Language Engineering throughout Europe.

10. References.

[ALMPR98] Gilles Adda, Josette Lecomte, Joseph Mariani, P. Paroubek, M. Rajman, "The GRACE French Part-of-Speech Tagging Evaluation Task", in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

[AK98] Adam Kilgarriff, "SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs", in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

[ATW96] Eric Atwell, "Comparative Evaluation of Grammatical Annotation Models", in Richard Sutcliffe, Heinz-Deltlev Koch, and Anne McElligott (eds), "Industrial Parsing of Technical Manuals", Amsterdam, Rodopi, 1996

[BLA94] E. Black, "A New Approach to Evaluating Broad-Coverage Parsers/Grammars of English", Proceedings of the International Conference on New Methods in Language Processing (NEMLAP'94), UMIST, Manchester, September 1994.

[DARPA99] Proceedings of the DARPA Broadcast News Workshop, February 28th-March 3rd 1999, Herdnon, Virginia, USA.