Project ref. no.:	LE4-8340
Project title:	Evaluation in Language and Speech Engineering (ELSE).
Deliverable status:	Public
Contractual date of delivery:	April 30th 1999
Actual date of delivery:	
Deliverable number:	D1.1
Deliverable title:	"A Blueprint for a General Infrastructure for Natural Language Processing Systems Evaluation Using Semi-Automatic Quantitative Black Box Approach in a Multilingual Environment."
Type:	Report
Status:	Final version 3.4
Number of pages:	62
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):	Editors: Marc Blasband & Patrick Paroubek. Contributors: Niels Ole Bernsen, Nicoletta Calzolari, Jean-Pierre Chanod, Khalid Choukri, Laila Dybkjær, Robert Gaizauskas, Steven Krauwer, Isabelle de Lamberterie, Joseph Mariani, Klaus Netter, Patrick Paroubek, Andrei Popescu-Belis, Martin Rajman, Antonio Zampolli
EC Project Officer:	Giovanni Battista Varile
Keywords:	EVALUATION, QUANTITATIVE, BLACK BOX, NATURAL LANGUAGE PROCESSING, MULTILINGUALITY, CONTROL TASK, PROPOSAL, FP5.
Abstract:	The ELSE project (Evaluation in Language and Speech Engineering) has for objective the study of a possible implementation of comparative evaluation in the field of Natural Language Processing in Europe. In the following document, we will present a general infrastructure for Language Engineering evaluation. This general infrastructure focuses on the definition of a (semi)-automatic and task-independent protocol framework for quantitative black-box evaluation of Natural Language Processing (NLP) systems in a multilingual environment. We will first present the concept of comparative evaluation as we see it, refining the distinction between the various kinds of evaluation that could be implemented for Language Processing. Then we will recall some background facts on the origins of the paradigm of evaluation and its current state of deployment in the field of Language Engineering both in United States and Europe. In the last part of the document we present a proposal for a deployment of comparative evaluation in Europe. We start by identifying the objectives for the various communities of actors, then we present the structure of an evaluation campaign and we propose candidate control tasks that could initiate the deployment of evaluation. Next, we address key issues related to the implementation of the proposed control tasks before concluding with resource considerations. In annex, we reflect on some practical considerations for the implementation of comparative evaluation.

Contents

1. Introduction

1.1 The ELSE Project.

The ELSE project (Evaluation in Language and Speech Engineering) is a Preparatory Action part of Language Engineering Program in the 4th Framework Program of the European Commission. Its aim is the study of a possible implementation of comparative evaluation in the field of Natural Language Processing in Europe.

In the following document, we will present a general infrastructure for Language Engineering evaluation. This general infrastructure focuses on the definition of a (semi)-automatic and task-independent protocol framework for quantitative black-box evaluation of Natural Language Processing (NLP) systems in a multilingual environment.

1.2 Preamble

1.2.1 What is Comparative Evaluation?

Comparative evaluation in language engineering has been used as a basic paradigm in the USA DARPA program on human language technology since 1984. Since then, other enterprises based on the same paradigm have been conducted in Europe, both at national and at European level, but on a smaller scale and over a limited time.

Comparative evaluation is a paradigm in which a set of participants compare the results of their systems using the same or similar control tasks and related data with metrics that are agreed upon. Usually this evaluation is performed in a number of successive evaluation campaigns with more complex data for every campaign and participation open from the start to anybody. For every campaign, the results are presented and compared in special workshops where the methods used by the participants are discussed and contrasted.

More precisely, Comparative evaluation consists of choosing or creating a control task, which may or may not correspond to the function of a complete system (such as a spoken language dialog system), or of a component (such as a morpho-syntactic tagger or a parser), of gathering system or component developers and integrators who are interested in testing their systems against those of other, of organizing an evaluation campaign which necessarily includes the distribution of linguistic data for training and testing the systems, and of defining the protocol and the metrics which will be used in the evaluation of the results. To give an overview of the past and present deployment of comparative evaluation, we list here most of the projects or programs that were a direct application of the paradigm:

- AMARYLLIS: a National evaluation campaign for Information Retrieval systems working on document written in French. First edition: 1996-1997, second edition: 1998-1999.http://www.inist.fr/accueil/profran.htm.
- ATIS (Air Travel Information System). DARPA Spoken Language Systems evaluation of spoken dialogue systems for air traffic database querying (1989-1995).

- GRACE (Grammaires et Ressources pour les Analyseurs de Corpus et leur Evaluation - CNRS): National evaluation campaign for morpho-syntactic taggers of French (1997) http://www.limsi.fr/TLP/grace.

- The Aupelf-Uref (International Association of French Speaking Universities) launched in 1994 the ARC program based on the evaluation paradigm for both spoken and written language for 7 different control tasks: Information Retrieval, Bi/Multilingual Corpus Alignment, Automated terminology Database Design, Message Understanding, Voice Dictation, Vocal Dialogue, Text-to-Speech Synthesis. http://www.limsi.fr/Recherche/FRANCIL/frcl.html.

- Morpholympics: evaluation of morphological analyzers for German (Germany, 1994)

- MUC (Message Understanding Conferences – DARPA): a long series of American evaluation campaigns for Information Extraction on several tasks related to the instantiation of given templates with information extracted from texts (published proceedings: MUC-3 to MUC-7 in 1998). http://www.muc.saic.com/.

- SENSEVAL/ROMANSEVAL (Evaluating Word Sense Disambiguation Systems for English and several Roman languages): co-sponsored by several EU-projects, ELSNET, ELRA and the British government, http://www.itri.brighton.ac.uk/events/senseval & http://www.lpl.univ-aix.fr/projects/romanseval

- SUMMAC (First Automatic Text Summarization – DARPA): three levels of summarization evaluation, user-based (May 1998). http://www.tipster.org/summcall.htm.

- TDT (Topic Detection and Tracking Project – NIST): segmentation, detection and tracking of a given subject in an information flow. First edition TDT-1:1997, second TDT-2: 1998 [DARPA99]. http://trec.nist.gov/tdt.html for (TDT-1) & http://www.nist.gov/speech/tdt98/tdt98.htm for (TDT-2).

- TEMAA (A Testbed Study of Evaluation Methodologies: Authoring Aids): European project, aimed at developing a user-based evaluation framework, following the EAGLES action. Application to spell checkers. http://www.cst.ku.dk/projects/temaa/temaa.html.

- TREC (Text REtrieval Conferences NIST and DARPA): a long series of evaluations in Information Retrieval including a main track and workshops on more prospective tasks like the recent addition of information retrieval on audio data. Proceedings published from TREC-1 to TREC-8 http://trec.nist.gov/.

Complementary to these project, the following projects address issues peripheral to evaluation related for instance to resources or tools:

1. DIET (Diagnostic and Evaluation Tools for Natural Language Applications): European project, aiming to develop methods and tools for glass box evaluation, for English, French and German. (began in 1997, successor of TSNLP) <>http://www.echo.lu/langeng/projects/diet/index.html.

2. FRACAS (A Framework for Computational Semantics): European project that has elaborated a set of 350 DQA tests (DQA: Declarative + Question + yes/no Answer) that illustrate (in English) almost 100 basic semantic phenomena. http://www.cogsci.ed.ac.uk/~fracas.

3. TSNLP (Test Suites for Natural Language Processing): European project aimed at developing systematic test suites to test syntactic capacities of NLP programs http://cl-www.dfki.uni-sb.de/tsnlp

4. MATE aims to facilitate re-use of language resources (spoken dialogue corpora at multiple linguistic levels) http://www2.echo.lu/langeng/projects/mate.

5. GATE Natural Language Processing and Computational Linguistics specific software architecture and development environment http://www.dcs.shef.ac.uk/research/groups/nlp/gate

And the next two projects deal with more generic aspects of evaluation like the methodology and infrastructure.

1. EAGLES (The Expert Advisory Group on Language Engineering Standards – Evaluation Workgroup): European initiative, one of which working groups has proposed a user-based methodology for evaluation. Two phases, EAGLES-I and since 1996 EAGLES-II. ,%20<A%20HREF=" ewg96.html? ewg96 projects issco-www.unige.ch http:http://www.ilc.pi.cnr.it/EAGLES/home.html</A>,%20<A%20HREF= (final report) & http://www.cst.ku.dk/projects/eagles2.html.

2. ELSE (Evaluation of Language and Speech Engineering): European initiative aiming to define a generic methodology for black-box, semi-automatic, quantitative evaluation <>& http://www.limsi.fr/TLP/ELSE.

The ELSE proposal differs from the USA DARPA evaluations in three ways:

- by trading competitive aspects of comparative evaluation for more contrastive and collaborative aspects through the use of multidimensional results and by putting a stronger emphasis on contrastive analysis of the methods and results.

- by being multilingual from the start.

- by considering usability criteria in the evaluation;

1.2.2 The Advantages of Comparative Evaluation.

The experience with comparative evaluation in the USA and in Europe has shown that the approach has significant advantages.

The performance of the system is not the major output of the evaluation exercise. More importantly, the evaluation will yield common knowledge for the participants and the funding agencies about the tasks, the metric and the techniques with which the problems can best be solved. The objectivity of the evaluation helps assessing the pros and cons of a technique. Note that being able to rerun the evaluation process with the evaluation data is a straightforward way to guaranty transparency and promote further use of the evaluation by-products.

For the stakeholders, evaluation has the following advantages:

- The comparative element provides a particular psychological incentive to the participants to deliver the best results possible.

- The developers benefit also indirectly from evaluation because complete evaluation

toolkits and by-product data become available afterwards.

- Institutions that have not participated in a campaign receive the possibility to evaluate their own technology in relation to the state of the art.
- At the same time, the paradigm of evaluation allows the funding agencies to measure if the money they have invested in technology development has led to significant progress and to identify areas where the technology needs further improvement.
- The commercial deployers and the end-users will be able to understand where the technology can help them and provide new solutions to the problems they face.

The USA DARPA example is very informative in this regard, as the results obtained for text dictation over the previous years show that it was possible to put dictation systems on the market. For more difficult tasks, such as unconstrained telephone dialogues, the poor results measured during evaluation campaigns show that more investment is still needed.

Resources.
A side effect of evaluation is often the production of high quality resources. Data are distributed to the participants in order to help them with training and testing their systems. As the participants need the data, there is an imperative to provide data of good quality and in due time.

The availability of metrics and measurement tools alongside with the data used for training and testing the systems, allows the participants to measure their progress. The data can be distributed in the community after the campaign is completed and re-used as training material in other campaigns.

Role of Evaluation.
Since 1984 DARPA has an up-to-date view of the state-of-the-art in language engineering worldwide through its series of evaluation campaigns. In front of the results, some non-US agencies now even request laboratories they fund do test their system in the Darpa framework. The experience of DARPA and others show that the comparative evaluation paradigm should be considered as a very powerful tool for research in the field of language engineering: the performance of the evaluated technologies and the understanding of the phenomena were significantly improved. The ELSE project thinks that this will remain an important factor in the future.

In language engineering a shift is taking place from theoretic and model-based approaches to more empirical and data-driven approaches. Systematic observation of corpora with real life speech and language becomes more and more important. The comparative evaluation paradigm fits naturally into this development.

2. Background

2.1 Why?

2.1.1 Why does Language Engineering need Comparative Evaluation?

The ELSE consortium has reached the conclusion that language engineering research that has the potential to lead to commercial success in the near future, is based on data. At this moment we have no applicable theory that allows us to deduce properties from first principles. Therefore, evaluation is required for validating hypotheses, for assessing progress and for choosing between alternatives.

The choice of criteria and metrics for the comparison is empirically deduced from the needs of the field and not from a theory. The successful usage of a metric is determined by the agreement of the actors in the field upon that choice. Many examples are needed to establish the key parameters of the metric. Comparative evaluation forces the agreement and provides the examples. The major success of the USA DARPA evaluation campaign, the recognition of the value of the HMM approach, is due to such an agreement on such a metric.

Furthermore, language engineering displays a paradoxical property. In many areas the state of the technology has reached a level barely sufficient to be usable in practice. Nevertheless, many commercial language-based applications do exist (e.g. machine translation, text summarization, dictation, spoken dialogue systems). Comparative evaluation could help resolve the issue, where the advertised performance claims are difficult to assess and compare objectively.

2.1.2 Why at European Level?

The major reason to have an international dimension is that the technology is international and multilingual. All the major developers and suppliers work on several languages, even if there are few real multilingual applications. Furthermore, all the major suppliers operate worldwide. Even if they do not take part in the evaluation campaign, they need an adequate infrastructure that comparative evaluation provides.

As the applications that are built for the end-users are often monolingual, one could argue that evaluation campaigns should be organized either nationally or in a linguistic region. The French national evaluation programs (e.g. GRACE and FRANCIL) have achieved positive results. Nevertheless, an international dimension is necessary to obtain the desired impact at the European level. Also the national research programs like VERBMOBIL in Germany or the NWO priority program in The Netherlands show very clearly that co-operation on European and international level is necessary for what concerns research.

Moreover, most European language markets are too small to allow for strictly national evaluation. A language with relatively few speakers (e.g. Danish, Dutch) can only rely on European co-operation to organize the evaluation campaign that they need. It is obvious that a language with no or inferior supporting computational tools will suffer in the competition between cultures. After film, television and music, the computer systems could become the next battlefield for the expansion or contraction of the European cultures. By organizing evaluation campaigns with an emphasis on multilinguality, the European Commission will support multiculturalism as it has done with the support of research programs and resources.

The porting of technology across languages becomes more and more prominent in the industrial field, leading to a greater need for multilingual comparative evaluation.

Finally, comparative evaluation is another way to make researchers of different countries

communicate and so forge a stronger European community for language research.

2.2 What Kind of Evaluation?

In analyzing the situation of evaluation, a clear terminology proved to be essential. In this section the ELSE project proposes definitions and key words that will be used throughout this report.

2.2.1 Concepts.

It is important to mention the differences that the ELSE project sees between competition, validation and evaluation in relation with the specification activities. The purpose of specification in this context is to determine before implementation the set of criteria used in the assessment activities and the reasons behind their choice.

- Competition uses only one criterion, a strict ranking of the participants based on that criterion, and little or no auditing of the participant's performances. As soon as more criteria are used, the competition becomes impractical, unless one participant is better than the others for all criteria. Sometimes a formula is used to compute one criterion out of several to create a competitive environment. But the choice of the formula is quite arbitrary and sometimes mathematically unsound. Competition is concerned with a local performance maximum.

- Validation uses a set of criteria derived from a detailed specification to audit the failures and successes. Validation is concerned with knowing whether a given performance threshold has been reached. Validation is usually performed for one product or one application with one set of requirements dependant on the intended usage.

- Evaluation uses more implicit criteria to perform the audit. The specifications are determined at the time of the evaluation. Evaluation is concerned with the reasons behind performance measure distribution and evolution. In evaluation, several technologies or systems are compared with a pre-defined set of criteria.

Depending on the situation, both validation and evaluation can be conducted on the same system.

Because many different criteria must play an important role, the ELSE consortium feels that the evaluations are multidisciplinary and that strict competitions are counter productive.

Every competition generates popularity and raises interest. The publication of the ranking of the participants in a comparative evaluation campaign has a similar effect, but it brings mainly short term benefits. Only the descriptions the participants make of their systems give an idea of the methods that were used to achieve the result. A system that performs well, because it has been hyped-up, is of much less interest than a system that does not perform as well, but shows a better conception or uses a promising, but not yet mature technology.

Also when the compared systems perform the same function, the differences in environment are such that any form of strict competition is meaningless. The ARISE project (http://www2.echo.lu/langeng/projects/arise/summary.html) showed this very clearly with four systems having the same function: automatically providing train schedule information by

telephone. The comparison of the different requirements and the different implementations is far more significant and important for the future of the field.

The difficulty of a competition is best exemplified by the comparative evaluation of translation systems. First of all, the quality of a translation is subjective. Secondly, it is not clear how to compare a cheap translator that gives a wrong grammatical output with an expensive one that is producing correct output. One can only validate the fit for their intended use through explicit criteria.

2.2.2 Different Types of Evaluation.

Different types of evaluation. Evaluation on a large scale is needed, but which kind? Looking at the whole development life-cycle of a technology, a few stages exist, each requiring the use of a particular type of evaluation. The ELSE consortium has identified the following five types (the first four are related to one stage, the fifth to all stages):

- Basic research evaluation tries to validate a new idea or to assess the amount of improvement it brings on older methods.
- Technology Evaluation tries to assess the performance and appropriateness of a technology for solving a problem that is well defined, simplified and abstracted.
- Usage Evaluation tries to assess the usability of a technology for solving a real problem in the field. It involves the end-users in the environment intended for the deployment of the system under test.
- Impact evaluation is the evaluation of the socio-economic consequences of a technology.
- Program evaluation can be seen as an attempt to determine how worthwhile a funding program (like LE) has been for a given technology.

User-oriented criteria are used in all five types when consideration of the end-user perception and behavior are included in the evaluation, e.g. the speed of speech, the acceptance of a mode of interaction.

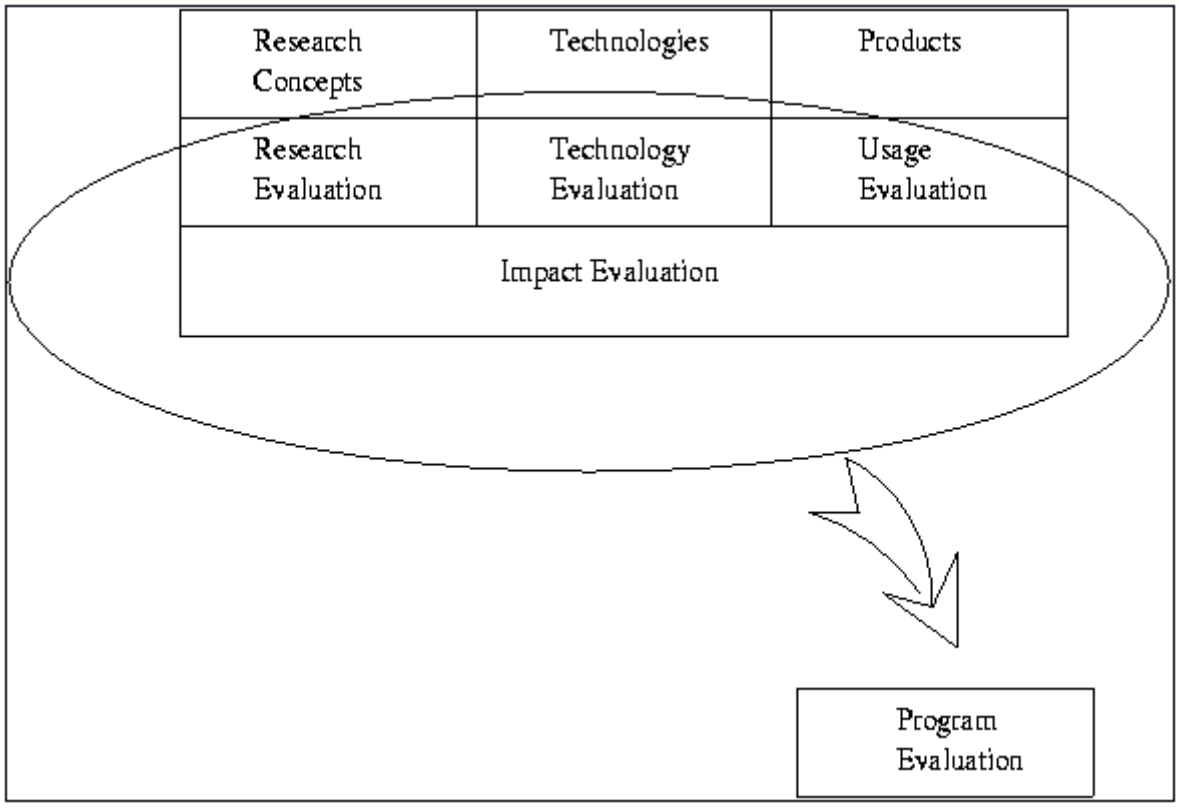

Figure 1

In the following chapters, we will only address Technology and Usage Evaluation. In the remainder of this chapter, we shall study the relation between the Technology and Usage Evaluation on the one hand and the Impact and Program Evaluation on the other hand.

2.2.3 Relationship between Basic Research Evaluation and Technology Evaluation.

Basic research evaluation can best be performed for new concepts that are expected to replace older ones. It tries to determine whether a concept is viable and if it provides a significant improvement for existing methods. When possible basic research evaluation can use previous results of Technology Evaluation to validate the fact that an improvement is brought under consideration by the novelty. Basic research evaluation can then measure:

- how large the difference with other state-of-the-art technologies is;
- how fast this gap is bridged;
- if these improvements have an asymptotic behavior.

It does not matter if the corpus and the metric are relatively old as the purpose of the basic research evaluation is to indicate a direction in an objective way.

2.2.4 Relationship between Technology Evaluation and Usage Evaluation.

In ELSE, we are interested in the deployment of comparative evaluation for Technology Evaluation and Usage Evaluation.

- Technology Evaluation measures the technical performance of a system used to perform a generic control task. The specific aspects of the application, environment, culture and language are abstracted from this task as much as possible.

- Usage Evaluation considers the actual performance of a system in the framework of a specific application, environment, culture and language. Not only technical aspects are compared but also usability criteria (in particular the human/machine interaction).

Thus, Technology Evaluation addresses a generic task. It must be abstract enough in order to provide a common reference ground for a sufficient number of participants to become interested and to accept to devote time and efforts for adapting their system to it. But it must also remain specific enough for allowing the participants to capitalize on their investment through the reuse in various (multilingual) applications of the specific developments they undertook and the knowledge they gained while participating to the evaluation.

Usage-Oriented evaluation considers the actual performance of a system in the framework of a specific application, for a specific language, and may result in the validation of the system, when compared with a priori requirements or user satisfaction criteria. Note that Usage evaluation is not necessarily comparative. The relationship between Technology Evaluation and Usage Evaluation is the following: if the technology is of insufficient quality, it will not fulfill the application requirements, while if the technology shows good results in laboratory conditions, it is not certain that it will obtain acceptable performances in the actual conditions of use.

Both the Technology and the Usage Evaluation use the results of a control task to perform the evaluation. Their main difference lies in the presence or absence of end-user considerations in this task: both try to establish how a technology performs the control task, but Usage Evaluation is also concerned with the usefulness of the task for the end-user and the related usability aspects for the systems under assessment.

The task used in Technology Evaluation is simplified and a number of its features are abstracted from the intended deployment environment. The key problem here is to abstract enough to get rid of the noise introduced in the measures by the specificity of the deployment environment while still remaining faithful to the real issues at stake for the progress of the underlying technology.

Usage Evaluation must take into consideration the attributes of the system that are essential for usability but not necessarily related to the technology itself: some measures then become irrelevant for the technology itself but are more related to ergonomic or even market related issues.

For a given application domain, significant problems still remain to be solved even after Technology Evaluation. These problems are not technological in their core. Because natural language is close to the human psyche, the behavior of the users and their reaction to the technology have a significant influence on the performance in actual field conditions.

In order to be reliable, Usage Evaluation must be exercised in real life situations with different environments, applications, languages and cultures.

The more one looks towards field aspects, the more the number of parameters to take into account increases, and the parameters themselves become more and more context specific. As we move from technology-oriented consideration towards field consideration, the complexity and size of the search space defined by the possible interactions of the input parameters increases drastically. One of the key results of the Usage Evaluation will be a reduction of this complexity with an understanding of which parameters have the largest impact on the users for what concerns usability.

The performance of a technology (as measured by the Technology Evaluation) tends to evolve through thresholds: an improvement of a technology inside an interval between two thresholds is not perceptible in Usage Evaluation. This interaction between technology and Usage Evaluation determines through these thresholds decision points for the industrial deployments of the technology: every new threshold that is reached defines a new class of applications that may successfully be deployed (marketability is another issue).

Comparative Technology Evaluation ignores the interface of the participating systems. But Usage Evaluation takes these interfaces into consideration. For Usage Evaluation, the measured performances and appreciation do not provide a clear distinction between the influences of the packaging of the system and of its core functionality. On this score, Technology Evaluation can be said to provide a glass box look on aspects that Usage Evaluation tends to handle in a black box oriented manner.

Ideally, Technology Evaluation should be able to predict some results of Usage Evaluation, because it takes place earlier in the lifecycle of a technology. But a large number of experiments are needed in earlier Usage Evaluations to show when these predictions are correct.

Technology Evaluation and Usage Evaluation are complementary. Both are needed as they each provide one part of the technical answer for assessing progress and for selecting a technology for a given application.

In the beginning of the lifecycle of a technology, one first expects to perform only Technology Evaluation, then Technology Evaluation and Usage Evaluation together, and when the technology has matured, only Usage Evaluation; until it is replaced by validation once a standard has been established.

2.2.5 Relationship between Technology Evaluation and Impact Evaluation.

The relationship between technology and impact evaluation is difficult to appreciate. These types of evaluation occur at distant points in the development lifecycle.

However, Usage Evaluation is closer in time to the full deployment of the technology. By the involvement of the end-users, it can predict the possible impact of the technology on the end-user as consumer or citizen. The relation however remains difficult and dangerous to make. It is only after several years that the socio-economic consequences which follow the recognition of an emerging technology can be fully appreciated.

Very often the prediction proves to be wrong, e.g. the paper consumption decrease expected from adopting computer technology for office work. The caution with which the main holders of speech recognition technology approached the market once the technology had passed the first trial of Technology Evaluation, is characteristic. They knew the market was there, they knew that the technology had reached a sufficient level of performance (at least for a single speaker in office conditions), but they also knew that the wrong market approach would kill the golden egged goose.

Depending on where we are located in the development lifecycle of a given technology, it is more appropriate to talk in the early stages of Impact Prospective Analysis and later on of Impact Assessment.

2.2.6 Relationship between Technology Evaluation and Program Evaluation.

As program evaluation contains a sum of all the other types of evaluation, the relationship between technology and Usage Evaluations and program evaluation is a straightforward one. Technology is one term of the sum, field usage is another and socio-economic impact a third term. They contribute one part of the general picture, namely the progress achieved by a given technology during the program. The progress can be quantified by several aspects of the results produced by the technology and Usage Evaluation, e.g. the performance improvement, the increase in the number of participants, the higher diversity of their origins, the augmentation of the number of languages handled for a given control task, the number of applications and environments where systems are deployed.

Naturally the relationship between progress and program quality is not linear. In conclusion, we could say that technology and Usage Evaluations provide some useful indicators for program evaluation, but not all of them.

2.3 A Bit of History

Some key points of the evaluation history of Technology Evaluation are presented (Usage Evaluation have never been performed in a comparative way except in one of the TREC tracks where the combination of the IR system and its operator is evaluated). This history helps understand the present state of affairs, both in the USA and in Europe. The ELSE project hopes to emulate the achieved results in the future.

2.3.1 The Evaluation of Speech in the USA.

In the 70 and early 80's, there was no standard measure or protocol for assessing the quality of speech recognition systems. Most technology developers claimed a 99%+ recognition rate, and several very different approaches coexisted in the research community, none offering a visible advantage over the others.

When DARPA started its new campaign in 1984, the Evaluation Paradigm (the comparative quantitative black box approach) was chosen as backbone for the program. Three years were required to design and implement the first evaluations which were run in 1987. The first formal tests showed that the general quality of the recognition systems was well below what was actually claimed, and that knowledge-based methods were bested by statistically based ones.

This allowed to resolve the issues and investment began to flow towards the methods which were identified as the ones that could lead to usable systems. In the mid 90's, they finally arrived on the market with big success.

The series of evaluation campaigns have demonstrated a global increase of performance (from separate word recognition to continuous speech recognition) in parallel with a full deployment of the technology that has been taken up by industry and now occupies a whole market sector. To give an idea of the progress made since 1987, it suffices to note that Philips, a firm that actively participated in the DARPA campaigns, is now, ten years later, advertising the FREESPEECH speech recognition system for 39 US$ on the Internet, and one of its competitors, Dragon Systems (also involved in the DARPA evaluations), offers a similar product at 49.99 US$ (August ’98 data).

2.3.2 Tools as By-products of the Technology Evaluation Campaigns.

Whenever comparative evaluation takes place, there is a need for deploying an evaluation software toolkit for validating the input data, for computing the performance results and for displaying them. The undertaking of such construction is often a costly enterprise, particularly when done on an individual basis. Comparative evaluation represents a strong factor of incentive for making such toolkit available to a given community. For instance, the SMART (ftp://ftp.cs.cornell.edu/pub/smart/smart.11.0.tar.Z), and its successors PRISE and ZPRISE (http://www-nlpir.nist.gov/works/papers/zp2/zp2.html), indexing tools produced during the TREC campaigns, have now become almost standard toolkits for information retrieval, while the SCLITE tool (http://www.itl.nist.gov/iaui/894.01/software.htm) for speech transcription comparison is widely used. These by-products of evaluations are a very important contribution for industrialists, who do not wish to devote specific resources to develop evaluation tools.

2.3.4 Subsidiary Control Tasks, Secondary Tasks, Hubs and Spokes.

As the campaigns progress, the control tasks evolve: subsidiary and secondary control tasks emerge. Refining a control task into optional subsidiary control tasks is something that generally develops over the years, as the issues associated with a control task become clearer. New problems, generally arising as satellite issues, are identified and come to the fore, sometimes replacing older ones as focal points of the evaluation.

For instance, in MUC-6 (1995) the sixth of the Message Understanding Conferences sponsored by DARPA, evaluation expanded in three new directions [LH98]:

- named entities (identification of persons, locations etc.);
- template elements (entities and their attributes);
- co-reference mark-up (linking multiple instances of the same entity).

This branching out of across the years of the main control task into various subsidiary control tasks is more clearly observed in series of American evaluation campaigns partly because of their duration. For instance, the speech evaluation series have developed a hub an spoke organization, each hub corresponding to different quality of audio signal, to which are attached different spokes (sub- control tasks). On other hand, the TREC Information Retrieval campaigns have seen the development of different parallel evaluation tracks, corresponding to specific aspects like for instance IR from audio databases (SDR), IR from very large sized databases, or lately the Question and Answer track where the distinction between classical Information Retrieval and Information Extraction becomes to blur.

2.3.4 Contrasts between USA and Europe.

For text related evaluation in the USA, the first MUC evaluation took place in 1987 and the TIPSTER program started in 1991. For speech processing, the first large-scale campaign (Continuous Speech Recognition and Large Vocabulary Continuous Recognition) also dates back to 1987. DARPA and NIST were the two funding agencies behind those campaigns. The American government provided an important budget that came along with evaluation objectives strongly inspired by military or political considerations.

The American campaigns have inspired similar efforts in Europe (e.g. SQALE). But the picture in Europe [JM99] is less homogeneous for several reasons. The amount of resources which has been devoted to evaluation until now is much less and comes from many different sources:

- The EU sponsors projects such as EAGLES, DiET, DISC, TSNLP, TEMAA, SPARKLE and MATE:
- In Germany, the Morpholympics (morphological analyzers evaluation) and VerbMobil (real-time speech translation) performed comparative evaluation;
- in France, the GRACE project (Part-Of-Speech taggers evaluation) sponsored by CNRS and the Francil ARCs sponsored by Aupelf-Uref);
- in the UK, the SENSEVAL/ROMANSEVAL co-sponsored by several EU-projects, ELSNET, ELRA and the British government.

Furthermore, the diversity of goals and infrastructures behind the different evaluation efforts in Europe is an extra factor adding to this heterogeneity.

It seems that the USA evaluation based programs had followed a top-down approach (the government strongly influenced the campaigns, but provided abundant funding and a long lasting infrastructure). In Europe, the strategy has been more a bottom-up one, with various efforts for which ELSE could be a converging point, leading to a more ambitious deployment of the paradigm of evaluation in FP5 as described in chapter 3 of this document.

2.3.5 Recent Examples of European Comparative Quantitative Black Box Evaluation Campaigns

2.3.5.1 The ARCs (Actions de Recherche Concertées) of the Aupelf-Uref

The International Association of French Speaking Universities launched the Francil research network on language engineering coordinated by J. Mariani, in 1994. It began in parallel with the ARC program based on the evaluation paradigm for both spoken and written language for 7 different control tasks organized as follows [JM98]:

Written Language Resources and System Evaluation (ILEC)

- ARC A1 Information Retrieval
- ARC A2 Bi/Multilingual Corpus Alignment
- ARC A3 Automated terminology Database Design
- ARC A4 Message Understanding

Spoken Language Resources and Evaluation (ILOR)

- ARC B1 Voice Dictation
- ARC B2 Vocal Dialogue
- ARC B3 Text-to-Speech Synthesis

The work for the first campaign started in July 1994 with the publication of a call for proposals which resulted in November 1994 in the selection of 50 proposals out of 89. An international advisory committee evaluates the program each year and includes experts of both ILEC and ILOR. Proposals were issued from 34 different laboratories.

The evaluation campaigns have a two year time span (1996-1997 and 1998-1999). Each control task has the same organizational structure, comprising an evaluator in charge of the management, a scientific committee whose members are the participants, one or more corpus providers and the participants. Except for ARC A4 (the number of replies to the call for proposals was not large enough to launch a complete program and resulted in a working group) all the ARCs were completed at the end of the first campaign, which was for some of them, an exploratory phase.

If we exclude A4, the total budget for the 6 ARCs was about 2 MEURO over 4 years, which averages roughly to 167 KEURO, per campaign, per control task (each with one evaluator and, on average, seven participants from three different countries) and for one language, French (A2 addressed French-English alignment). Note that only the evaluators and the corpus provider were funded. The participants only received a token subsidy to cover a part of the cost of adapting their system to the test conditions and the travel expenses they incurred.

All the campaigns used quantitative black box evaluation metrics except for ARC A3, for which qualitative assessment by domain experts was used (evaluation metrics for B2 are still being defined but it will very likely use a metric inspired by the PARADISE framework [WLKA97][WHFFM97]).The results of the first campaign were presented and discussed at a series of workshops organized as satellite events of the Journées Scientifiques et Techniques du Réseau FRANCIL in April 1997.

Some essential information about the first campaign:

ARC	# Participants + # Evaluators	Countries	Approximate Corpus Size	Metrics
A1	8 + 1	CA, CH, FR, USA	330,000 documents,. indexed by 28 topics	Precision & Recall
A2	6 + 1	CA, CH, FR, UK	2.8M words aligned FR & Eng	Precision & Recall at various levels of granularity
A3	8 + 1	CA, FR	3,800 journal pages indexed + thesauri	Qualitative assessment by domain experts
B1	5 + 1	CA, FR	100 hours by 120 speakers + 40 M words text corpus + 64 K words phon. lexicon 4 Language Models (approx. tot. 170K words)	Word Error Rate (NIST/Sclite V3.0)
B2	5 + 1	CA, FR	30 hours of dialogue	under development
B3	9 + 1	B, CA, CH, FR,	2,100 sentences (27.3 K words)	Phoneme Error Rate (modified NIST/Sclite)

The first evaluation campaign resulted in:

- a better knowledge of the existing systems in each area, as well as a better assessment of the state of development of the domain;
- precise evaluation metrics defined in collaboration with the participants;
- well documented and tested language resources;
- the creation in each domain of a community of actors interested in evaluation.

2.3.5.2 GRACE (Grammars and Resources for Analyzers of Corpora and their Evaluation)

Started upon the initiative of Joseph Mariani from Limsi-CNRS and Robert Martin from INaLF-CNRS, GRACE was part of the French program CCIIL (Cognition, Intelligent Communication and Language Engineering), jointly promoted by the Engineering Sciences and Human Sciences departments of the CNRS.

Initially GRACE [ALMPR98] was intended to run over four years (1994-1997) in two phases the first dedicated to Part-of-Speech tagging for French text, and the second, which has since been abandoned, was intended to tackle syntactic analysis, also for French. The first year was devoted to bootstrapping the program by:

- installing the different committees which compose the organization (a co-ordination committee in charge of management, a scientific committee, and the participants);
- preparing the call for tenders;
- identifying and obtaining the linguistic resources;
- specifying a first version of the evaluation metrics.

The call for tenders was published in November 1995. The training corpus was distributed globally to all the participants in January 1996, while the dry run corpus was distributed individually to each participant in an encrypted form during the fall of 1996. The results were discussed during a workshop restricted to the participants, a satellite event to the Journées Scientifiques et Techniques du Réseau FRANCIL, in April 1997 [ALMPR97]. The test corpus was distributed in the same manner as for the dry run, at the end of December 1997. The preliminary results of the tests were discussed with the participants in a workshop in May 1998. The final results were disclosed on the WEB during fall of 1998 as soon as they had been validated by the organizers (cross validation with two different processing chains based on different algorithms and developed at two different sites) and the participants.

At the beginning there were 18 participants from 5 different countries (CA, USA, D, CH, FR), from both public research and industry, and 3 evaluators (Martin Rajman, at first from ENST then EPFL, had joined the initial members of the coordinating committee who were from INaLF and Limsi). The 2 corpus providers were also the initial evaluators (Limsi and INaLF).
Out of the 21 initial participants, 17 only took part in the dry run and only 13 completed the tests.

The size of the training corpus was around 10 million words and it consisted of texts evenly distributed between literary works and newspaper articles. For the dry run, the participants tagged

a corpus of roughly 450,000 words with a similar genre distribution and the performance measure was computed over 20,000 words to which a reference description had been manually assigned. For the tests, the participants had to mark a corpus of 650,000 words and the measure was taken over 40,000 words.

The real cost of GRACE is difficult to estimate because:

- the program was interrupted for a short period;
- none of the evaluators worked full-time on the project;
- the evaluators were located at different sites;
- some of the evaluators were so enthusiastic that they gave some of their own time to the project.

Nevertheless, it is possible to give the following assessment. Over the 4.75 years that the project lasted, the travel and consumable expenses can be roughly estimated at 100 KEURO. A minimal estimate of the evaluator work is of one person working full-time during the whole project. If we assume a yearly overall cost of 150 KEURO, we come up with a total in the order of 800 KEURO over 4.75 years. In GRACE only the evaluators were funded (the participants were only reimbursed their travel expenses) and the previous cost does not include any cost for the data as the corpus providers were the evaluators themselves.

We can estimate that a participant that followed the project from the beginning, contributed a minimum of 2 person/weeks. If we compute the cost over a two years period, we find a total of 335 KEURO for one control task, one language, 3 evaluators and 13 participants from 5 different countries. Note that this cost is double the estimated cost of one ARC campaign whose characteristic numbers are half of those of GRACE, but it would be hasty to infer a linear relationship between the cost of a program and the number of participants from such scarce data.

GRACE used the quantitative black box metrics: Decision and Precision, which were derived especially for GRACE from the metrics used in Information Retrieval (Precision and Recall). One of the lessons to draw from the GRACE experience, is that ideally, results should be cross-validated with two different processing chains, based on different algorithms (when this is possible) and developed at two different sites in order to ensure their accuracy and quality.

Project	# Participants + # Evaluators	Countries	Approximate Corpus Size	Metrics
GRACE	13 + 3	CA, USA, D, CH, FR	10M words + 60K words hand tagged + 350K word lexicon	Precision & Decision (an adaptation of I.R. Precision & Recall)

The results of GRACE are:

- better knowledge of the existing systems in each domain and of their state of development;
- precise evaluation metrics defined in collaboration with the participants;
- a new product on the market, as one participant decided to add a tagger to his catalogue as a result of his participation in GRACE;

- the creation of a community of actors interested in evaluation;
- a new linguistic resource which did not exist previously, i.e. the 1,000,000 word corpus in French with Part-Of-Speech tags tested, obtained by comparing and validating the output produced by the participants (the object of the MULTITAG project, funded by CNRS in the Ingénierie des Langues program, sequel to CCIIL).

2.3.5.3 SENSEVAL/ROMANSEVAL (Word Sense Disambiguators Evaluation)

SENSEVAL [AK98] is a pilot evaluation campaign for Word Sense Disambiguating systems [IV98] working in English. It was coordinated by Adam Kilgarrif (who kindly provided the cost information below) and was run in collaboration and in parallel with the ROMANSEVAL evaluation campaign, the same exercise as SENSEVAL but applied to the French and Italian languages. ROMANSEVAL was coordinated by Jean Véronis (LPL-University of Aix-en-Provence), Frédérique Segond (XRCE-Grenoble) and Nicoletta Calzollari (CPR-Pisa).

The SENSEVAL exercise proposed two distinct tasks: one for those who need sense-tagged training data, and one for those who do not. For both, tagging was only performed on a few selected words, which were supposed to be tagged with the senses defined by HECTOR (both a dictionary and the associated corpus). HECTOR is the result of a collaboration between Oxford University Press and Digital. Initially, a third task had been proposed, for systems not requiring training data and which would be asked to tag all the words of the test material (using WordNet [CF98] senses). However, it was abandoned, partly because the resources to support it were lacking. Note that the texts to be tagged were excerpts and not full documents. There was no distribution of untagged corpus material of the same genre as that to be used for evaluation, but the evaluation material was taken from a similar spread of genres to that found in the British National Corpus.

SENSEVAL/ROMANSEVAL ran over 8 months, from December 1997, when the first expressions of interest were registered, to the final workshop in September 1998 in Herstmonceux (UK).

The dry run data samples were distributed in March 1998 to the participants, who had to return a formal agreement to participate, which included the license for research use of the HECTOR data (copyright Oxford University Press, which provided the data for free). The dry run data consisted of a sense-tagged mini-corpus of 40 word types: 20 nouns, 10 adjectives and 10 verbs, all the HECTOR instances (e.g. between 300 and 1000) of each type, as well as the HECTOR dictionary definitions for these 40 word types and another 200 for the porting of programs which take Machine Readable Dictionaries as input.

Test training data (word samples and lexical entries) were distributed in June 1998. The tests were done on 20 nouns, 20 adjectives and 20 verbs. At this stage, legitimate activities included developing, maybe semi-automatically, the algorithm-specific lexical representations for the target words, or manually identifying sense-mappings between HECTOR and another resource (e.g. WordNet [CF98]).

In early July, the participants were asked to freeze their software and the test data for all tasks were distributed. The taggings were returned during the first half of August. The results were made available to the participants at the end of August and disclosed at the final workshop in September 1998. Note that the participants were given up to mid-October 1998 to improve their score if they wished to do so, provided they did not modify their system. This opportunity gave them the chance to correct spurious errors (like the one due to format discrepancies for instance)

or to optimize the learning of their system.

Initially, about 35 teams claimed interest in participating in SENSEVAL, and in the end, the results of the evaluation of 21 systems (including derived versions) were presented at the final workshop, along with the results obtained with different baseline approaches.

A very rough estimate of the cost of the pilot-SENSEVAL (which dealt only with one language, English) is the following:

	KEURO
Coordinator gross salary (6 person-months)	23
Overheads on coordinator salary (approx. 45%)	11
English manual tagging: grant from UK EPSRC	16
English manual tagging: support in kind from Cambridge University Press	3
English lexicon and corpus, provided free by Oxford University Press	0
Results computation (paid in kind by paying travel and workshop attendance)	1
Student assistants (paid in kind by paying travel and workshop attendance)	2
Hardware and computing	0
Workshop: venue hire	5
Workshop: printing, photocopying, workshop subsidies	2
Total	61

Note that the biggest chunk is the coordinator salary (data, hardware and computing were provided for free), and a lot depends on how well-paid, and how efficient, the coordinator is. The participants were not funded. According to the organizers, they were constrained in task definition by the availability of resources, particularly the dictionary. Cost estimates are not available for ROMANSEVAL, but the data were provided by ELRA at a very low cost and small off-the-shelf electronic dictionaries were used.

Both SENSEVAL and ROMANSEVAL used quantitative black box metrics. Concerning metrics, SENSEVAL marks an important milestone, since it was during the final workshop that the use of a cross entropy measure in conjunction with a penalty value (based on sense hierarchy distance or functional communicative distance between the correct and the proposed sense for a token) as proposed in [RY97] was recognized to be more relevant than the metrics generally used in the literature (Boolean Tag Error Rates), because of their higher discriminating power.

Project	# Participants + # Evaluators	Countries	Approximate Corpus Size	Metrics
SENSEVAL	21+1	FR, USA, IT, UK, CH, KO, MA, CA, SP, NL	60 lemmas in 8,448 contexts	Weighted Cross Entropy
ROMANSEVAL	7+1	FR, IT, CH	60 lemmas in 3,724	Precision/Recall

			contexts	per form/lemma pair

The results of SENSEVAL/ROMANSEVAL are:

- better knowledge of the existing systems and of their state of development;
- the acceptance of a new evaluation metric by the community;
- the experience gained in the management of a collaborative multilingual evaluation pilot project involving 3 languages (there will be a special issue of Computers and the Humanities on SENSEVAL, edited by Martha Palmer and Adam Kilgarrif);
- annotated data;
- the creation of a community of actors interested in evaluation (specific working groups have been organized, and planning for a meeting about the next campaign has already been arranged).

2.3.6 Lessons from History.

The most important lesson is that Technology Evaluation of the technology can bring significant results by identifying the most promising technology and by showing the rate with which it improves.

The impetus that evaluation gives to all the participants generates better systems, fosters agreements on metrics and measures, and insures the existence of substantial corpora. The tools generated as by-products of the campaigns provided significant support to the field.

In the past, the campaigns were successful when the technology had the potential to be improved. Some campaigns showed that the technology had reached a ceiling and that further evaluation would not show enough improvement.

It also became obvious that in order to be successful, a campaign needed the co-operation of the participants and agreements on all major points: protocols, metrics and measures as well as data.

Proposing an organization that refines the control task along various specific dimensions in a non-contractual manner for the participants, is a good way to broaden the scope of a campaign and to prepare the next evaluation campaign. It also contributes to refining the picture that is drawn of the current state of a given technology.

Evaluations were successful in the USA because a strong political and governmental support and involvement drove the process and financed a large part of it. It is not obvious whether the same political support can be obtained in Europe.

2.4 Criticism

The evaluation campaigns, specially the campaigns performed in the USA, have been criticized on two major points: the reduction of innovative ideas and the choice of the control task.

Innovative Ideas.
Some critics argue that the comparative evaluation can hinder the development of newer

technologies because by necessity newer technologies will provide worse results before showing their value.

This can be solved by using evaluation campaigns of longer duration (two years seem to be a good compromise) and by funding long-term research projects with meeting points that are located far enough in the future to allow for the development of new ideas.

The ELSE project proposes to separately handle basic research with a specific evaluation mechanism, working with longer term objectives. In addition, Technology evaluation results and by-products will also support the development and evaluation of new ideas.

Other critics argue that comparative evaluation may also break off innovative ideas because the focus is put on a single approach and only short term considerations will be taken into account by the researchers. This possible problem will tend to occur when a strong competitive element is introduced in the comparative evaluation. The project ELSE however, attempts to reduce the competitive element of the comparative evaluation.

Finally, we think it is not comparative evaluation that kills new ideas, but the use one makes of comparative evaluation. The same could be said of any other scientific tool. To convince oneself of this, it suffices to look back on the dampening effect that the first publication in 1969 of Marvin L. Minsky and Seymour A. Papert book entitled "Perceptrons" [MP90] had on the field of neural network computing (this excellent book addressed the limitation of one type of connectionist model which people unfortunately generalized to all connectionist models). The problem lay more in the way we conduct science than in evaluation.

The Control Task.
A last criticism is related to the choice of the control task used for comparison. It may not necessarily be related to the key function that is needed to determine the practical value of the technologies. Using the evaluation paradigm is like using a very powerful lamp to help finding an object in a dark room, but focusing the light on a place where the object is not. A counter argument would be that not using the evaluation paradigm is like trying to find this object without a lamp. Moreover, Usage Evaluation would tell in which corner of the room the lamp should be directed.

3. Proposal

3.1 The Objectives of Evaluation.

With this proposal, the ELSE project attempts to build upon the result already obtained in the past by comparative evaluation and to incorporate new demands from the fields while reaching for a wider audience. The underlying factors that could inspire a large scale evaluation effort in Europe are based mainly on scientific but most of all, on economic grounds.

3.1.1 The Objectives of Evaluation for the Developers.

The comparative element provides a particular psychological incentive to the participants to deliver the best results possible.

After each evaluation, a workshop will be held in which the participants will explain their analysis of the task and the techniques with which they have solved it. This knowledge will be shared. The performance of the system is not the major output of the evaluation exercise. More importantly, the common metric used and the knowledge gained during the evaluation will be shared by the participants and by the funding agencies during workshops.

It may happen that better results are obtained by some participants because they have used data of better quality. Therefore, evaluation will help identify better data, not only better techniques. It also contributes to assess the impact that the quality of data has on system performance.

The objective evaluation will advantageously complements paper publications by weighting scientific ideas using real data common to all the participants. The results reported in papers have sometimes been obtained on specific data, or with specific measures, that are hard to generalize and do not always meet the common evaluation requirements of the metric.

The developers will also benefit indirectly from evaluation because complete evaluation toolkits and by-product data become available after the completion of an evaluation campaign. Institutions that have not participated in a campaign can evaluate their own technology in relation to the state of the art by using the resource of that completed campaign.

Another important by-product of the evaluation campaigns is the broad agreement about metrics and measures.

3.1.2 The Objectives of Evaluation for the Community at Large.

The evaluations allow the funding agencies to determine if their investment has led to significant progress. They will help identify areas where the technology needs further improvement.

The commercial deployers and the end-users will be able to understand where the technology can help them and provide new solutions to the problems they face. The full evaluation program will also provide indications of the applicability of the technology to practical solutions and show the importance of the technology for the society at large.

Evaluation is a way to identify promising technology and to show its value to industry, thus speeding up the time required for a concept to become a mass-market product.

3.1.3 Language Engineering's Current Need for Data.

Language engineering needs resources to progress. Right now it is obvious that, for languages other than English, there is a lack of:

1. Part-Of-Speech tagged corpora and treebanks;
2. ontologies;
3. lexicons;
4. corpora tagged with word senses (taken from a reference dictionary);
5. large corpora of speech transcriptions (aligned with voice data).

Evaluation provides a partial solution to the problem through the production of standardized,

annotated and validated linguistic resources at a low cost, from the data processed by the participants during an evaluation campaign. Evaluation contributes to the definition of the associated standards with a practical viewpoint. An essential asset, because without standards the resources would not be usable.

Of course, evaluation can directly contribute to the support of evaluation itself. Once the best technology has been identified in a given domain, it can be harnessed to produce annotated data for future evaluation campaigns.

3.1.4 The Contribution of Usage Evaluation.

The ELSE project proposes to complement Technology Evaluation with Usage Evaluation in order to take also into account more market oriented considerations. The specific goals that the ELSE consortium think Usage Evaluation should achieve are listed below.

Usage Evaluation will clearly show the value of a technology for the user and will allow to measure progress in this direction over several campaigns.

Usage Evaluation will provide clear directions to the choice of criteria for Technology Evaluation. Indeed, this choice is empirical and should be determined by the usage of the technology and not by the technology itself. Usage Evaluation will clarify the relation between technology and usage by answering a number of questions:

- Which technology is required for a given usage? In which case is a good user interface preferable to an average technology and where is the technology indispensable?
- Which metrics and criteria should the Technology Evaluation use to obtain the best correlation with Usage Evaluation?
- What are the thresholds for the technology that are indispensable for a specific usage?

The Usage Evaluation that the ELSE project proposes will break new grounds, because of its innovative aspect. It will answer questions like:

- Among the many parameters describing the real world environment which are the one with the greatest influence on the evaluation results?
- How can judgement measure be performed so that they are stable, repetitive and reliable?
- Which metrics are relevant for the usage of language technology?
- What is the influence of language and culture on the usability of the technology?

Because it is closer to the usage of the technology, Usage Evaluation will support its commercial deployment and show which are the key elements that influence the successful deployments.

3.2 Structure of a Campaign.

Within one campaign, the same control task is performed by all participants. Their results are the

basis for the comparisons.

To define these control tasks, ELSE proposes to use the **abstract architecture** of a generic application that covers all the aspects that language technology needs to address nowadays.

The generic application we will use as a reference frame for defining specific evaluations, will be a cross-language intelligent information extraction system. Here 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. We will use the architecture of this generic application for communicating and explaining the relationship between the various evaluation tasks that we intend to propose, each evaluation task corresponding to an abstract functionality or module of the architecture. The various components of the architecture will be developed along the following 3 activities which corresponds to different segments of the loop that language information would follow when user would interact with the generic application (from the user to the application, then back to the user):

1. **Information Profiling** (data analysis, e.g. input speech signal transcription).
2. **Information Querying** (dialog management issues and mapping results to query, e.g. document selection).
3. **Information Presentation** (output modality selection, language generation, e.g. speech synthesis).

Evaluation points can be selected at the input and output of individual modules of such architecture and also at any point along arbitrary chains of modules. Thus, new evaluation tasks can be defined by linking various modules of the abstract architecture in a braided fashion [KNRGRC95]. The global functionality achieved by the so selected evaluation chain will define the control task to be used for Technology Evaluation. The characteristics (usability requirement) of the Usage Evaluation that could be performed to complement the Technology Evaluation would be drawn from the characteristics of the environment existing at the end points of the chain of module. Of course, these environment parameters are quite different depending on whether one sees the chain of module as an embedded module in a larger system, or whether one sees it as a stand-alone application. In the latter case, the parameters are more numerous in order to cover additional ergonomic issues.

Our abstract architecture is very much like the new DARPA COMMUNICATOR development/evaluation paradigm [COM98], where a real information software (derived from the JUPITER system [JPSSJGTH98]. developed at MIT), will be distributed to all the participants for module development or improvement. However in our case the architecture is not a real one, but an abstract one and is used only as a reference framework for linking the various evaluation tasks.

3.2.1 The Control Task.

A control task is the function that the participating systems perform during evaluation with the conditions under which this function must be performed (e.g. for parser evaluation a control task could be bracketing of the constituents in texts of minimal size). The common evaluation protocol will use quantitative black box metrics deployed around the control task.

In addition, we put the following generic requirements on the definition of the control task:

- It should be easy to communicate to and be learnt by non-specialists;
- A human operator should be able to perform it (at least in terms of function but not necessarily with the same performance level, particularly in term of speed);
- It is not necessarily identified with the functionality of a specific component in any classical NLP system architecture;
- A common reference formalism must be agreed upon. It must have sufficient expressive power to describe the processing performed in the control task.
- The principles behind the metrics used to compare the performance of the systems with their usability, should be straightforward and easy to communicate (unfortunately, this does not imply anything about their implementation).

3.2.2 Baseline and Metrics.

For each control task, a baseline performance level may be determined either by straightforward implementation of a basic algorithm or based on economical considerations (e.g. currently, for optical character recognition the economical threshold is 99.7% error free performance, below this level, it is cheaper to resort to keyboarding). Because it is representative of the state of the field and of the difficulty of the task, we think that a baseline approach should always be part of the results to provide a contrastive point of view over the systems' performance.

Sometimes human intervention in results quality assessment cannot be avoided. But whatever the assessment procedure, evaluation result production should be automated as much as possible, in order do be easily reproducible (a guaranty for transparency of the protocol). Several different tester should be used, as they will not always agree, even if they apply the same evaluation criteria (inter-tester agreement statistics like the "kappa" on [COH60][COH68][KRI80], are extremely useful for this, e.g. see its use in ROMANSEVAL [VER98]). Note, that the less technical a control task is, the harder it is to provide for it a reliable metric as quality criteria and performance measures tend to be based on subjective value scales (e.g. text summarization, translation, speech synthesis). In general, Usage Evaluation require many more subjective parameters, and better quality results are obtained when the testers are taken among the intended end-user population. Production of evaluation results should be automated as much as possible, in order to be easily reproducible, a guarantee for transparency.

3.2.3 Basic Requirements for Evaluation Data.

Linguistic resources are needed to build and annotate the reference data set used to compare the participating systems. For some tasks, the resources may already exist, but are unsuitable for comparative evaluation because it is very likely that the potential participants will already have had access to the material (such material can only be considered as training material and not as reference material).

Concerning the annotations themselves, human intervention will necessarily be required. Otherwise it would mean, that the task could be properly carried out by an automation and that evaluation will then be unnecessary. This last remark puts the focus on one of the reasons behind the high cost of evaluation for language engineering. To build the reference data set, we need human intervention. These data are used for comparing data produced by automatic means. As

computer capabilities keep on increasing in terms of speed and the amount of data handled, comparing several automatic approaches requires more and more data in order to exhibit significant differences. Despite the help that can be brought by dedicated software, the rate at which humans can produce data is almost constant because of inherent biological limitations (e.g. there is a maximum rate at which one can transcribe audio data).

A way to limit human intervention in building reference data sets consists of:

- distributing a large amount of data for processing by the participants;
- only evaluating on a piece of that data.

The size of the data sampled for evaluation is then defined by two contradictory requirements. It should be:

- as small as possible in order to limit the production cost;
- big enough to contain as many different language phenomena as possible and allow the production of significant differences in the evaluation results.

3.3 What?

3.3.1 Six Candidate Control Tasks for Technology Evaluation.

The main areas of language engineering that are current central preoccupations of researchers and developers, are [MLIM98]:

- dialogue management;
- translation;
- information retrieval/extraction.

The RTD priorities for Human Language Technology in FP5 listed in [HLT98] are:

- adding multilinguality to systems at all stages of the information cycle;
- enhancing the natural interactivity and usability of systems;
- enabling active assimilation and use of digital content, through language.

Considering the current state of the domain, we have identified the following 31 possible candidate control tasks which could be easily specialized into subsidiary tasks. We use (T) for Text, (S) for Speech, and (G) for Generic i.e. pertaining to both Text and Speech application domains. Where an evaluation has already been performed, an example is provided with the name of the exercise, the sponsor, its nationality and the year.

1	G	Language Models	[ARC B1/Aupelf/FR/95]
2	G	Translation Memories (sub-sentence level matching and partial clause analysis)	.
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	Lemmatisers	[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	

The next table presents the list of previous tasks according to the three activities identified by the ELSE abstract architecture: Profiling, Querying and Presentation.

	Information Profiling	Information Querying	Information Presentation

Speech	Speech Recognition	Database Dialogue Querying	Speech Synthesis
	Speech Segmenting		
	Coreference Resolution	Information Extraction	
	Named Entities Extraction		
	Topic Detection & Tracking	Multilingual Data Alignment	
	Language Identification		
	Speaker Verification	Machine Translation	
	Language Understanding		
Generic	Language Models	Translation Memories	
Text	Text Segmenting	Machine Translation	Text Generation
	Part-Of-Speech tagging		
	Lemmatising		
	Predicate Argument Structure	Multilingual Data Alignment	
	Parsing		
	Named Entity Extraction		
	Word Sense Disambiguation	Information Extraction	
	Coreference Resolution		
	Topic Detection & Tracking		
	Language Understanding	Summary Generation	
	Hand Writing Recognition		

Out of the previous control tasks we have pre-selected the following six. They could be used for the first Technology Evaluation campaigns. We would like to see this list validated by the actors of the domain in general and in particular by key representatives of the language engineering industrial sector.

The criteria for selecting these tasks were:

- evaluation is possible;
- the by-products of the evaluation can be reused.

Finer selection criteria will have to be applied when implementing these control tasks. The criteria should at least be based on the potential number of participating systems and on the linguistic resources available when starting the evaluation campaign.

1. Broadcast News Transcription; [DARPA99]

2. Cross-Lingual Information Retrieval / Extraction [GG97];
3. Text To Speech Synthesis [RS97];
4. Text Summarization [MITR98];
5. Language Model Evaluation. (Word Prediction Task) [JARD98];
6. All or a selection of the following techniques: Part-Of-Speech tagging [ALMPR98], Lemmatization [HAUS94], Analysis of Syntactic Functional Relations [BLA94] [ATW96], Word Sense Disambiguation [AK98].

Comparing these three tasks with the researchers preoccupations and the priorities of FP5 give the following two tables. These points of interest are reasonably well addressed.

	Multilinguality	Interactivity	Digital Content
Broadcast News	X		X
Cross-Lingual Information Retrieval / Extraction	X	X	X
Text To Speech Synthesis		X	X
Text Summarization	X		X
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.

3.3.2 Data Resources.

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 tags, Lemmas, Syntactic annotations and Word Sense tags.

The evaluation of these six tasks will produce data resources (see table above). Out of the 30 reusability possibilities for these resources between the six control tasks, 17 are actually possible. Each time, the data produced by one evaluation is interesting in the scope of another evaluation. If the reuse of data can take place between two control tasks, it is important to remember that such an operation entails scheduling constraints for the two evaluation campaigns. The second evaluation cannot start until the data produced by the first evaluation have been completely processed.

Producer/Consumer	BNT	CLIR	TTS	SUMZ	LM	TECH
Broadcast News	Reuse	Reuse		Reuse	Reuse	Reuse
Cross-Lingual Information Retrieval / Extraction		Reuse				
Text To Speech Synthesis			Reuse			
Text Summarization		Reuse	Reuse	Reuse	Reuse	Reuse
Language Model Evaluation		Reuse	Reuse	Reuse	Reuse	Reuse
Technique		Reuse	Reuse	Reuse	Reuse	Reuse

It is obvious that the data of one campaign can be used to start the following one for the same control task (diagonal of the previous table) .

3.3.3 Complementary Usage Evaluation.

Related to four of the six control tasks for Technology Evaluation, meaningful stand-alone control task for Usage EvaluationEvaluation could be the following ones:

Control Tasks for Technology Evaluation	*Control Tasks for Usage Evaluation*
Broadcast News	Transcription of Virtual Meetings
Cross-Lingual Information Retrieval / Extraction	Multimodal tourist information
Text To Speech Synthesis	Text-to-speech for the blind
Text Summarization	Text summarization of financial newspapers
Language Model Evaluation	
Technique	

These tasks were chosen with the knowledge that applications or prototypes exist. A major difficulty will be to find the participants for these stand-alone control tasks for Usage Evaluation.

If these cannot be found, one should think about the value of these stand-alone tasks: it may not represent any real demand, but the possibility exists however that the developers of the systems do not wish to participate.

For the two remaining control task (Language Model Evaluation and Technique), specific Usage 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.

As in other cases it is expected to evaluate the performance of deployed systems with real users and real demands.

In all cases multidimensional comparisons are necessary to cover the complexity of the tasks. Comparisons with several criteria are proposed (but not competition with a unique criterion).

To complement Technology Evaluation, other stand-alone control tasks can be thought of for the Usage-Oriented evaluation, the previous are given as indications of what could be done.

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

Instead of these six control tasks, one task can be proposed that covers the whole spectrum of activities: news on demand. This task searches multimedia material for information that is relevant for a given query. The purpose of news on demand is to provide archived broadcast news material. A query is formulated and video excerpts from past material are extracted from archive databases.

News on demand encompasses the major research directions that were identified:

- Dialogue management to handle the details of the query and the navigation in the space of the response;
- Translation of the query to search in material annotated in another language;
- Information retrieval and extraction by the definition of the task, where the images provide an additional, multimodal dimension to the task.

This control task will also show how the priorities of FP5 are worked out in the field:

- Multilinguality by using material in another language than the language of the request;

Natural interactivity to handle the details of the query and the navigation in the space of the response;

- Use of digital content of the stored information.

Finally this control task is itself a stand-alone control task for Usage Evaluation and thus allows to perform Usage Evaluation the latter in a very natural way.

3.4 How?

Within the objectives of the project ELSE, the imbedding of evaluation in the practical organization of the research projects of HLT had to be looked at. The project recognizes two modes of operations:

- Proactive: where the control tasks are decided beforehand (in particular before participants selection), e.g. as defined in the previous section;
- Reactive: where the tasks depend on the projects that are chosen from every call for proposals of HLT and on their clustering.

The proactive evaluation can be organized and prepared before the proposals of the first call of FP5 are accepted. The reactive evaluation campaigns can only be started when the accepted proposals are known and the clustering of projects are completed.

3.4.1 Clustering Considerations.

Project clustering was initiated in the course of FP4 in order to contribute to the objectives of the program, and to the achievement of the performance criteria laid down for the Telematics

Application Program [JD95] (as FP5 is not very explicit on clustering, the data of FP4 is used here). The purposes of the program were:

- adopting a concerted approach;
- exchanging information and experience;
- developing the use of standards and best practice;
- optimizing the reuse of resources across project boundaries.

In a broader perspective, project clustering was also meant to support the long-term objectives of the LE sector, which are mostly motivated by market considerations:

- to increase awareness in the business community of the potential of LE technology;
- to promote the LE sector in the marketplace;
- to establish end-user oriented goals on behalf of the sector as a whole;
- to maintain a watch on developing a market for LE products and services;
- to provide market feedback to the projects;
- to increase sensitivity to market requirements amongst the projects;
- to assist in optimizing productivity of RTD;
- to promote quality throughout the sector;
- to increase awareness across clusters.

Possible factors which have been considered to organize the clustering of projects, are [LGLK98]:

- application delivered to the user as the result of the project;
- language technology used within the project;
- technological function achieved through the use of LE technologies;
- market;
- market opportunity.

Of these five factors, market opportunity was identified as the most appropriate basis for clustering. Grouping by technology had been termed to be "very attractive" in terms of project cross-fertilization and quality improvement. However, it was deemed impractical because of the large number of technology combinations, and not rewarding enough as concerns the target user community [JD95].

Evaluation can provide an important inter-project and inter-cluster link and an exchange medium that would contribute to the objectives of the program and the long-term objectives of the section. It will bring the projects of a cluster together in one common activity, which will force more inter-project communications.

Once a control task has been drafted after identifying a need for validation in a cluster, a careful

selection of the features of the control task should be done in order to allow the largest number of systems to participate. As the purpose is comparison and not competition, the metrics should only be defined to measure what the application performs and should not be calibrated to provide fairness between projects or even clusters.

3.4.2 Multilingualism.

For a given evaluation campaign, the problem we face here is the following: How to compare the solutions proposed by N different systems for a given problem in M different languages. Performing an evaluation for all possible NxM combination between languages and systems is no realistic because such undertaking would imply the production of the data required for evaluation for all the M languages and to port each of the systems to the other M-1 languages.

The idea is to reduce the number of languages to process without losing the significance of the evaluation results.

The ELSE consortium has identified two means reduce the number of languages one need to consider to run an evaluation:

- cross language functionality requirements (specifying different input and output languages);
- using a pivotal language where a system functions in at least two languages, the one it was initially designed to work with, and the second which will be common to all the evaluated systems (this scheme was used in the SQALE [SYLL97] project);

The generalization of the evaluation results to languages different from the ones which are used in the control task, could be helped if detailed comparative information about language specific features and their correspondence across languages was available (e.g. in the language lineage as French, Spanish, Italian, Portuguese and Romanian all derive from Latin).

The cross-language requirement scheme has the advantage of a more flexible architecture and avoids the problem of choosing a pivotal language. The drawback of this scheme is that it always requires an extra functionality akin to translation. The translation may not be part of the initial functionality under test, when the task is basically monolingual. It may also significantly increase the noise in the evaluation measurements. Furthermore, cross-lingual requirements cannot be successfully applied to tasks which are intrinsically monolingual like speech recognition, speech synthesis, or lemmatization.

The key issue is to find a way to distinguish methodological aspects (which are generic across languages) from the linguistic knowledge (which is specific to a particular language). The evaluation protocol could require that all the participating systems separate these kinds of information (e.g. using a clear-cut distinction between architecture, program and linguistic data). This would still not be sufficient to compare the different programs and is not always practical from an implementation viewpoint.

For Usage Evaluation, no solution exists for the Multilinguality other than a careful selection across different languages of the application characteristics, the end-users population types, the deployment environment specificity and the usability requirement analysis (particularly performance).

3.4.3 Phases of Evaluation.

An appropriate duration for the completion of an evaluation campaign seems to be two years. Less, the participants do not have the time to capitalize on the results of the previous campaign; more, their motivation might falter. When starting the first campaign of a new evaluation program, it would be wise to plan for a preliminary one year period devoted to advertisement, control task awareness build-up, community establishment, metrics preliminary definition and data selection.

Almost all the American evaluation campaigns have followed the organization described below, which some refer to as the paradigm of evaluation. This is also true for most of the previous European efforts based on a quantitative black-box evaluation protocol.

Ideally, the running of an evaluation campaign should comprise four phases:

Phase 1 - Training

- The potential participants and the audience targeted by the control task are identified. The control task details are advertised. The potential participants are queried for their interest.
- At this stage, the activities linked with Usage Evaluation are the following: the specifications are reviewed, standardized and compared. The end-users are identified. The mechanism to collect the results is determined, insuring they will be representative, comparable and reproducible. The survey and questionnaires are defined.
- The criteria and metrics for comparing the systems are proposed as well as a standard to present the results.
- The training data is collected, formatted and distributed to the participants who formally agreed to participate in the campaign. The data are not necessarily annotated data or complete training data. They need to be representative of the data that will be used for the tests, and must cover a large enough set of intra-domain variations. The best way is to set aside a portion of the training data for the three phases separately: training, dry run and test.
- The protocol is communicated to the participants as are all the formats and support tools that will be needed to exchange data during the evaluation campaign.
- Training data distribution can either be restricted to the participants or made public. During this phase, the evaluator is responsible for building the dry run reference data which will be used in the next phase and for documenting or refining the rules specifying the gold standard.
- Communication of the gold standard definition can happen here, but may be postponed until the beginning of the dry run phase.

Phase 2 - Dry run

- Real but non-public run of the evaluation protocol with full involvement of the participants on the dry run data. The size of the dry run corpus and the time constraint imposed on the participants make manual processing impossible. A participant should not be able to identify the portions of the corpus which will be used for performance measurement. The dry run data are sent encrypted and the key is communicated

simultaneously to all the participants who are allowed a limited time to process the data. Synchrony between the participants is important to ensure fairness and transparency of the protocol.

- For the Usage Evaluation part, the participants present trial applications with specifications, criteria, presentation of the results and collection mechanisms written according to the agreed standards. These are compared, corrected and improved.

- This phase is concluded by a final judgement on the dry run data, the measures and the gold standard, during which the feedback of the participants is taken into account. At the same time, the participants are formally asked to choose to have their results disclosed or not, answering either way is required to participate to the next phase.

Phase 3 - Tests

- The test phase is a rerun of the dry run phase on new data, but the judgement concerns only the test data and the reference data.

- Every participant in the complementary Usage Evaluation performs his own evaluation. The results should answer two major questions: How does the deployed product fit the requirements, once it is in use? Do the requirements adequately predict the population of users of the system and their behavior?

- The test phase ends with a workshop in which attendance is restricted to the participants. They are asked to present their systems or algorithms and comment on the relationship between their results and the methods they used. Signature of a non-disclosure agreement of the workshop contents is mandatory to participate. This phase ends with the publication of the results and distribution of the training data, as well as the other by-products of the evaluation (mostly annotated and validated data).

Phase 4 - Impact study

- This phase deals with assessing the benefits brought by the evaluation campaign to the field: the increase in amount of annotated and validated data, the identification of promising directions and new algorithms, new products whose creation was a consequence of the evaluation campaign, new actors and progress made by the participants.

- The comparison in the Usage Evaluation will help measure the influence of the different usability criteria on the results and so determine which criteria are important and which are not. The comparisons will also help in determining which threshold values are needed for the various technologies used in the tests to be usable in practice.

- The results of this phase can be used to modify the protocol for future campaigns.

3.4.5 Results Computation: A Model for Coherence Verification of Quality Measures used for Natural Language Processing Systems Evaluation *(by Andrei Popescu-Belis)*.

3.4.5.1 Introduction.

Natural language processing (NLP) uses two modalities of scientific investigation: explanation and construction. The former aims at understanding the language phenomenon, while the latter builds models and implements systems in order to process some aspects of linguistic utterances. Whether the systems are supposed to corroborate descriptive theories, or serve as technological applications, it is absolutely necessary to prove, using validation and evaluation protocols, that the systems have indeed the expected properties. Evaluation is thus an estimation of a computer system utility with respect to a given task and a given category of users. In this section, we locate Evaluation in the broader frame of software engineering, in respect of specification and verification. We outline a formal framework for automatic, result-oriented evaluation which divides the process of evaluation into three phases: *measure* (measuring the software capacity to perform a certain kind of processing), *rating* (assessing the quality of each measured result) and *assessment* (integrating the assessments of the various capacities). This model is then used to present coherence criteria for numeric measures: *upper/lower limit*, *indulgence/strictness*, and *monotony*. Application of the framework and of the criteria is then illustrated by two examples.

Evaluation: Software Engineering and NLP.
To get a more accurate view of evaluation, we shall situate it in the broader frame of software engineering. Evaluation is thus the moment, in the development cycle of a project, when the result is compared to the initial goal, and its utility for the project beneficiary is measured [MS58]. When the goal of the project is the realization of a system or a piece of software, evaluation takes place at the end of the programming cycle, as stated by software engineering theory [SOM92][HAB97]. The programming cycle is traditionally divided into three stages: *specification*, *realization* and "*verification & validation*". Verification only makes sense with respect to well-defined specifications. Software specification is a very rich domain, which has evolved from using natural language, to schematic representations, and then to specification languages such as VDM, Z or SADT [LG90][HAB93].

The specification of a given task must be as precise as possible, remaining however independent from any particular solution of the task. So, specifications must be: non ambiguous, complete, verifiable, coherent, modifiable, reusable. Formal specification languages, as well as means to prove specification well-formedness, have been developed. Verification, in its turn, comprises various control operations [ILL90]. *Verification*, strictly speaking, means verification of a software conformity to its specifications. *Performance check* means the measure, and often comparison, of qualities which are not contained in the formal specifications, e.g., speed, user-friendliness, need of various resources, etc. As for verification, programs may be analyzed without execution (static analysis or program proving), but they have also to be tested through execution (operating tests); one can also test conformity to various standards. Passing the first two tests means *validation* of a system, while passing the last one is *certification*.

Evaluation versus Verification.
So, what does evaluation mean for NLP software? NLP is generally considered as a branch of artificial intelligence because there are no known algorithmic solutions to the problems it addresses [LAU87]. With respect to specification, there are two kinds of artificial intelligence tasks. Some tasks allow for a formal specification of the desired state: the program thus possesses the sufficient means to recognize by itself that it has reached the desired state, as in for instance in problem solving, 'automated theorem proving' or various games like chess or checkers.

On the contrary, NLP, or for that matter pattern recognition, are subject to two levels of specification. The *form* of the desired solution to a given problem can be formally specified: for

instance, a system may be designed to tag each word in the input text with a label obeying precise syntax (as in morpho-syntactic analysis), or to classify input patterns into one of the predefined categories (pattern recognition).

However, there seems to be no formal method to specify that a solution is the *correct* one. In the two examples above, neither the label, nor the category, can be recognized by the program itself as the correct ones, as there are no formal methods to define correctness. The human designer defines the correct answers using examples and implicit knowledge (e.g., linguistic competence or categorization capabilities), which have not yet been explicitly formalized. Thus, the goal of e*valuation* in NLP is to measure the extent to which a program satisfies the non formalizable part of the specifications. As such, evaluation is closer to 'performance check' than to'verification, as described by software engineering. The theoretic framework presented below models such a process.

Black Box, Glass Box and Modularity

A was said before, there are two broad types of verification in software engineering: black box and glass box [ILL90]. In the former approach, the test data is chosen only according to the specified relations between input and output, without considering the internal structure of the tested software. The latter approach, however, takes into account this structure in order to build more precise test data; thus, one can focus on testing the more sensitive parts of a program, and possibly compute the test coverage rate (the proportion of the program code effectively verified).

Both approaches are useful to NLP evaluation (beyond of course *verification* of NLP software as any other piece of software). However, glass box evaluation sometimes means considering directly a program structure or knowledge, without running the program on any test data. Also claiming to be glass box, some approaches have produced generic test data (or test suites) without any consideration for the structure of the program under test, but based on a variety of linguistic phenomena which are unavoidable in NLP tasks. Black box evaluation in NLP keeps its original meaning, i.e. evaluation of a system result for a given input.

Choosing one of the approaches depends of course on the evaluation purpose. For instance, the black box approach is frequent in comparative evaluations. The glass box approach is more analytical, and has often been promoted by European generic evaluation projects. One should add to these approaches the user-based methods, which measure the satisfaction of human users (experts or not) with certain capacities of the evaluated system. Of course, this approach seems particularly straightforward in NLP, where the best criteria for language processing assessment is human appreciation, but it is difficult to conciliate with the requirements of comparative, large scale evaluation in a multilingual environment. When the task on which systems are evaluated can be divided in several sub-tasks, these may be evaluated separately in order to obtain more detailed information on each processing stage. However, not all the systems use the same division in sub-tasks. Also, modular evaluation is more costly, as it needs test data for each stage or module. Indeed, it is controlled data prepared by the evaluators that should be fed into each module of the processing chain, and not the output of the previous modules, in order not to penalize the modules which are preceded by poorer modules.

Another possibility for modular evaluation is to replace, in an existing system, one module with the corresponding one from another system, and estimate the variation of quality. Thus, modular evaluation is also a way to evaluate components which are not useful by themselves to the human user. In what follows, the meaning of "evaluation" will be limited to system evaluation. According to [KSJ95], system evaluation has three phases: evaluation of a system proximity to a given goal, evaluation of a system relevance to a given task, and failure diagnosis. This analysis somehow seems both incomplete and redundant, and the ISO 9126 [ISO91] framework will be preferred here. This standard provides guidelines for software quality assessment. *Quality* has six

characteristics: functionality, usability, efficiency, reliability, user-friendliness, maintenance possibilities and portability. Of course, these apply also to NLP software; however, given the distance between the goals of NLP and its current capacities, evaluation focuses generally on the main *functionality*. Sometimes user-friendliness and efficiency are also evaluated.

3.4.5.2 A Formal Framework for Quality Measures

We have already outlined the relations between verification (of a program conformity to its specifications) and evaluation (conformity to the non formal aspects of specifications). Evaluation enables one to know whether a given method or piece of knowledge contributes effectively to solve a problem, or to accomplish a task that satisfies a given user; evaluation also allows one to compare different methods or pieces of knowledge. Numeric measures seem to bring a convenient exactitude to the answer.

At a more general level, the estimation of a system *quality* by a single number resembles the economic concept of *utility*, related to the idea of *price*. Of course, research policies sometimes have to make exclusive choices, and thus need single scale comparisons between various systems and approaches. The EAGLES report [MK96] states that: "evaluation is a function relating objects and users to something we will call utility. Utilities can sometimes be expressed in financial terms and represent a consistent preference relation among the items utilities are assigned to."

The directives of the ISO 9126 [ISO91] standard divide the evaluation process in three stages: quality requirements definition, evaluation preparation, and evaluation procedure. The second stage is divided in three phases (metrics selection, rating levels definition, assessment criteria definition), which match the three phases of the third stage (measurement, rating and assessment). The first division seems however too general, and designed merely for the organizers of evaluation campaigns. So, the evaluation model proposed below elaborates merely on the division of the second stage.

The Three Stages of Evaluation
According to the ISO 9126 standard, the "quality" of a piece of software can be decomposed in attributes, which are grouped into the six characteristics cited above (functionality, usability, efficiency, reliability, user-friendliness, maintenance possibilities and portability). In order to evaluate a system on a given task, for each attribute a *metrics* shall be defined, which yields a score on a given scale. These scores are then transformed into marks or rating levels. Finally, several rating levels have to be combined if one wants to provide a single result for each system: a single score is less informative, but more adapted to comparative evaluation.

More formally, let S be a system for which several *capacities* (or functions) have to be evaluated, say C_1, C_2,…, C_n. These elementary capacities are, ideally, disjoint. First, each capacity has to be estimated using a measure (or metric) *mu* , which provides a value on a given scale, say [0; 1] (other intervals can be easily converted to this one). Such a measure depends on the measured capacity, and is not linked (at this stage) with any kind of utility.

Measure *mu*: $\{C_i \mid C_i$ is a capacity of $S\}$-> [0; 1]
C -> *mu* (C) = V_C (value of capacity C of S)

Formula 1

Each measured value has then to be appreciated with respect to the desired values, and thus

transformed into a *score*, which may be continuous or discrete (a set of "marks", $\{n_1, n_2,\ldots, n_p\}$). The appreciation (or rating) phase thus depends on the objectives of evaluation, the evaluators' needs, the state of the art, etc. and provides a judgment of each capacity. Of course, each capacity is appreciated according to specific criteria.

Appreciation *alpha* $_C$: [0; 1] -> $\{n_1, n_2,\ldots, n_p\}$ or [*min*; *max*]
V_C -> *alpha* $_C(V_C) = A_C$ (appreciation of the capacity *C*)

(Formula 2)

Finally, one has to summarize the various appreciations in order to produce a final result, in case this kind of abstraction is sought, i.e. if evaluation has to produce a single numeric score. Actually, a set of appreciations A_C provide more information than a single score, and the EAGLES methodology [MK96] considers the set as the final result – leaving to other evaluators the task of combining the A_C .

Summarization *beta*: $\{n_1, n_2,\ldots, n_p\}^k$ -> $\{n_1, n_2,\ldots, n_p\}$
$(A_{C1},\ldots, A_{Ck})$ -> *beta* $(A_{C1},\ldots, A_{Ck}) = A_S$ (result of the system *S*)

(Formula 3)

In order to simplify notations, the user has not been explicitly taken into account here, and the fact that each phase is also conditioned by the user needs is implicit. According to the more general notation of the EAGLES guidelines, evaluation is a function of the following type: *E*: *Systems x Users -> Values*.

Besides, in real evaluation campaigns, a phase can be shunned from the three phases above (often the 2nd is implicitly incorporated in the 1st), or the order can be reversed (for instance the 3rd may become 2nd). Also, the scale (or appreciation set) may vary, but this is only a matter of conventions, and remains fully compatible with our framework.

Definition of the *mu* Measure.
It is only very seldom that one can measure directly a capacity, i.e. without running the system. For instance, one can measure a dictionary size or a number of rules, but these are quite trivial "capacities". Most of the interesting capacities have to be measured by running the system on some input data, here linguistic input, e.g. textual data sets. One of the evaluation duties is to estimate, from results on several input data sets, the "real" value *mu* $(C)=V_C$ of the respective capacity *C*.

More formally, let *C* be the capacity to perform some processing on input data D_i belonging to a certain class *DELTA* , *DELTA* = $\{D_1, D_2,\ldots, D_n,\ldots\}$; the D_i are very often natural language texts, so *DELTA* is numerable, but non-finite, and very large. We note $m(D_i)$ the measure of a system capacity to process the input data D_i. Conceptually, *mu* (*C*) should be an "average" of *all* the values $m(D_i)$, the results on all possible input data, for instance their arithmetic mean.

Theoretic measure: *mu* $(C) = FI$ [$m(D_1), m(D_2),\ldots, m(D_n),\ldots$], *FI* is a "mean".

(Formula 4)

The only effective way to compute *mu* (*C*) is to approximate it using a subset of the possible data *DELTA*, say *DELTA* $_{test} = \{D_1, D_2,\ldots, D_k\}$. The system capacity to process this data (a finite and

small subset) is measured using the system response on each input data, noted $rep(D_i)$. So, the theoretical measure *mu* is replaced in reality with a measure *m* that computes the quality of a response: $m(rep(D_i))$. Averaged on the test set $DELTA_{test}$ (using the same mean *FI*) this measure is considered to estimate *mu* (*C*). The choice of $DELTA_{test}$ is a key factor for the evaluation objectivity. One may attempt to build a $DELTA_{test}$ that characterizes the whole *DELTA*, or possibly only part of it, in order to evaluate the system for various sub-domains, or with limit cases.

Effective measure: $mu\ (C) \sim FI\ [\ m(rep(D_1)), m(rep(D_2)),\ldots, m(rep(D_k))\ ]$, *k* small.

(Formula 5)

This reduction of *mu* (*C*) to an average of $m(rep(D_i))$ makes the evaluation problem more tangible, but does not really *solve* it. In order to specify the protocol of an evaluation task, and especially the way scores are computed, one has to measure the system ability to process a given input data D_i, i.e. its proximity to (one of) the expected answers, as computed by experts.

Description of the *m* measure.
In order to define the computable coherence criteria given below, the analysis will be limited here to the systems for which one can compute a quality measure of the response with a given input data $m(rep(D_i))$. It is not possible to compute $m(rep(D_i))$ in a deterministic way, otherwise the system task would be itself deterministic, which is obviously not the case for NLP problems. It will thus be required that $m(rep(D_i))$ may be computed automatically using the correct or expected response (the "gold standard"), which is defined for the test data by human experts. The scoring method should obey the following framework:

a. it is possible to describe the test data $D_1, D_2, \ldots, D_n$;

b. with the help of experts, it is possible to build one or several correct responses $K(D_i)$ for each test data D_i. This set is the *key* or *gold standard*. The quality of the experts' agreement on the key may be measured using the *kappa* measure [COH60][COH68][KRI80];

c. it is possible to describe the set of all possible responses, noted $REP(D_i)$. Then, $K(D_i)$ *is included in* $REP(D_i)$ and of course $rep(D_i)$ *belongs to* $REP(D_i)$ where $rep(D_i)$ is the system response. Let $REP = Union_i\{REP(D_i)\}$;

d. a *quality measure* is a (computable) function *m*, where *m*: *REP* -> [0; 1], which measures the quality of the response $rep(D_i)$, or its proximity to the key. So, $m(rep(D_i))$ is the system success on D_i. This value may or may not be transformed, using another function, to obtain a "mark" (in the appreciation phase).

If there exists a distance *d* on the response set $REP(D_i)$, then a quality measure may be defined, based on this distance, by:

$m(rep(D_i)) = min\{d(rep(D_i), k(D_i)) \mid k\ belongs\ to K(D_i)\}$, or
$m(rep(D_i)) = d(rep(D_i), k(D_i))$ if the key set $K(D_i)$ contains a unique key $k(D_i)$.

(Formula 6)

However, it is quite infrequent to find a distance on a complex space such as $REP(D_i)$. It should be noted that the property $d(r1,r2) = 0 \Rightarrow r1 = r2$ (compulsory for *d* to be a genuine distance) is not really necessary, because two different answers may be appreciated at the same level $m(rep(D_i))$.

Summarizing the appreciations
The appreciation phase, or transformation of the measured value into a note or mark, has to be adapted depending on the desired notes or marks. A common constraint is that these be comprised in the [0; 1] interval, where 0 qualifies the worst responses, and 1 the best (i.e. correct) ones.

The different marks obtained by the various capacities of a system are combined into a single score (if such a synthesis is needed) in the summarization phase. Of course, not all the measured capacities may be relevant to the final score: this depends on the evaluation purposes. When a single score is needed to summarize the evaluation, comparison between scores is used to compare systems. The problem is that, unlike the [0; 1] interval which is strictly ordered, the set of appreciations (before summarization) is not well ordered – whether it is $[0; 1]^n$ or $\mathbf{R}^n$, where **R** is the numeric axis. Several methods may be used to transform the $[0; 1]^n$ set into [0, 1]: consider the lowest score, or compute an average score (arithmetic or harmonic mean).

3.4.5.3 Meta-evaluation of quality measures

According to the definition, $m_C(rep_S(D))$ is the measure of the capacity *C* of a system *S* to process input data *D*, or, in other words, a measure of the response quality. An interpretation of "quality" will be first examined, then coherence criteria will for such measures will be expressed.

Objective quality
There are at least two epistemological perspectives on $m_C()$. The first one considers the measure as a neutral operation, dictated by the capacity *C*. The idea of quality is nothing but subjective, and becomes apparent only through the appreciation phase, which links the scores to their effective "values" for the user.

The second perspective believes that "quality" is an objective property of a system (with respect to a given task and user), a property that could, at least in theory, be measured by asking a lot of experts about the system. The measure plus the appreciation are designed to make visible, automatically, this objective quality; in other words, they associate to the "hidden" capacity levels some numbers or labels that make them visible.

Considering the combination of measure and appreciation phases, this establishes a mapping between a set of objective quality values, manifested in English by ‘perfect’, ‘good’, ‘average’, ‘poor’, ‘zero’ (horizontal axis) and the score interval [0; 1] (vertical axis, see figure 2). If the scores belong in fact to a discrete set of labels, then the measure plus appreciation computes the labels for the "real" quality levels. However, it will be admitted here that the scores or quality measures belong to the [0; 1] interval (or 0–100%), because a discrete set makes notations less clear, the model less general, and evaluation less sensitive. Discrete labels could better be used afterwards, for the summarization phase.

Regarding the following criteria, a distinction has to be made between the objective quality set (horizontal axis, figure 2) and the set of scores (vertical axis). As the first set is not directly accessible, strictly speaking, the criteria expressed using this set (or axis) are subject to *argumentation*, whereas those expressed using the second axis are subject to mathematical *proofs*.

Both kinds of criteria have to be satisfied, by the measures, for all possible test data. One may in this case find counterexamples, that is, show that a measure does not fulfill a certain criterion, for some input data D_i . These are acceptable if D_i really belongs to the subset of *DELTA* considered for the evaluation. For a counterexample, D_i may be a very specific text, specially constructed to

show the measure problem.

Coherence criteria for quality measures.

Upper limit (UL) condition

The first condition is quite obvious: when a response *rep*(*D*) is 'perfect' (from the experts' point of view), then the quality measure should yield the maximal score (100%). Indeed, it would be odd if the maximal score was never attained. But the reciprocal condition should also hold: the maximal score should be attained only by perfect answers, i.e. belonging to the key set. As the key set is precisely known (this being an hypothesis of our framework) the (UL) criterion can be proved for a given measure *m*.

UL: $m(rep(D)) = 100\% <=> rep(D)$ *belongs toK*(D)

(Formula 7)

Lower limit (LL) condition

Conversely, there is a coherence criterion for the opposite end of the axis. The worst responses (most poor ones) should receive à 0% score, and reciprocally. Unlike the (UL) criterion, the set of the most poor responses is not well delimited, as it has little interest for the task; its interest comes only from evaluation. So:

LL: $m(rep(D)) = 0\% <=> rep(D)$ is one of the worst responses

(Formula 8)

The (LL) condition is only subject to argumentation, because it depends on the conventional definition of the worst responses; let $REP_W(D)$ be this set.

The first term of (formula 8) can be rewritten using the following equivalence $m(rep(D)) = 0\% <=> rep(D)$ *belongs to* $m^{-1}(0\%)$, which expresses that the response belongs to the set of responses which receive à 0% score, i.e. $m^{-1}(0\%)$. The (LL) criterion requests that the two sets $m^{-1}(0\%)$ and $REP_W(D)$ be identical, so it can be divided in two paired criteria. First, the measure should verify

LL-1: $m^{-1}(0\%)$ *is included in* $REP_W(D)$ (subject to argumentation)

(Formula 9)

In other words, all the responses receiving a 0% score should indeed be among the worst possible. It is generally uneasy to describe theoretically all the responses noted 0%, and it is impossible to circumscribe $REP_W(D)$, so this criterion is seldom examined, for instance only using some examples.

Conversely, a measure should also obey the (LL-2) criterion, i.e. all the worst responses are indeed attributed à 0% score.

LL-2: $m^{-1}(0\%)$ *contains* $REP_W(D)$ (subject to argumentation)

(Formula 10)

This criterion is easier to study (despite not being subject to proofs) using concrete examples of 'very poor' responses. It is quite easy to find examples of poor responses: for instance, no processing at all of the input data, random or trivial processing, etc. The (LL-2) criterion requests that these responses receive a 0% score, or at least very low ones (in the LL-2´ version below). If

discrete scores are used, these examples of poor answers should receive the lowest label.

The idea behind this criterion is that if a measure always yields scores above 50%, then a score of 60% does not have the same meaning as the same 60% for a measure that really starts at 0%. So, a coherent quality measure should cover the whole interval [0; 1]. In its non formal version, the criterion reads:

LL-2´: *all the poor responses should receive low scores*

(Formula 11)

A question that the criterion (LL-2´) depends on is: are there any responses rated 0%? If there are none, then necessarily (LL-2) cannot be satisfied, unless there is no poor response (from the experts' point of view). This is the justification of the following sub-criterion, which is subject to mathematical proof:

BI-3: $m^{-1}(0\%)$ *is different fromØ*

(Formula 12)

The corresponding non formal version is:

BI-3´: *the minimal possible scores have to be low*

(Formula 13)

These criteria may be difficult to study when the quality measure is the combination (e.g., summarization) of several measures that evaluate the same response *rep*(*D*), for instance the *f-mesure* of recall and precision scores. In order to verify the lower limit criteria (LL), it is not enough to check that each measure reaches a 0% (or low) score, but they have to do this for the same responses. Otherwise, the final result of summarization can have a lower limit much above 0%.

Indulgence/severity of two measures

To choose a particular evaluation measure, it is often necessary to compare it with other possible measures. When, for the same response, a measure always yields scores greater than those of another one, the first measure is said to be *more indulgent* than the second one (and the second is more *severe*).

IS: m_1 is more indulgent than m_2 iff
for all D, for allrep(*D*),
$m_1(rep(D)) = m_2(rep(D))$ *or*
$m_1(rep(D)) > m_2(rep(D))$

(Formula 14)

This makes provable the idea of "indulgent rating". Proofs are not however easy to find, as the comparison must hold for all the input data *D belongs toDELTA* . It might also happen that a measure m_1 is more indulgent than m_2 for some subsets of *DELTA*, and more severe for other subsets. The point in comparing the indulgence of two measures is to be able to choose the most sensitive one given the expected results of the tested system(s): if the results are good, then a severe measure may be chosen, but if they are poor, then an indulgent measure is more adapted. In this way, the differences between scores are maximized.

The notions of absolute indulgence or severity are more difficult to justify: they would make sense if an "objective" measure could be found. One may substitute to such a measure some examples of responses with their "objective" scores, as fixed by experts (for instance the worst responses, as in (LL-2), or the best ones). When several measures are compared (for the same capacity), the most indulgent one could be "absolutely" called indulgent (and the same for the most severe). Of course, this supposes that evaluators are able to compare all measures.

Regularity of a measure

Regularity is the property of a measure to yield greater scores when the responses get better according to the experts. The measure is then uniformly increasing; this is of course a highly desirable property.

REG: m is regular iff *for all* D, *for all* $rep_1(D)$, *for all* $rep_2(D)$,
$[rep_1(D)$ "better than" $rep_2(D)] =>$
$[m(rep_1(D)) = m(rep_2(D))]$ *or* $[m(rep_1(D)) > m(rep_2(D))]$

(Formula 15)

This condition is not subject to proof, because "better than" is not a computable relation. This relation may be computed locally, for response that differ very little. But it may be easier to find counterexamples, and thus show that a given measure is incoherent, by choosing two responses that are trivially ordered (one "better than" the other) and finding out whether the scores really reflect this order.

Graphical synthesis

The criteria for evaluating quality measures (i.e. for meta-evaluation) have been enounced using common sense arguments, familiar to those who have studied evaluation closely. The compliance of a measure with these criteria can sometimes be proved, or only asserted using arguments such as examples. Besides, counterexamples may also prove that a measure does not fulfill certain conditions.

A graphic representation of the possible biases of an evaluation measure is given figure 2. This presupposes that the response quality may be represented on a linear scale that goes from ‘very poor’ or ‘null’ to ‘the best’. Thus the ideally objective measure transforms this scale linearly (uniformly) into the score interval [0%; 100%]. Figure 2 represents the relative positions of various measures (the precise values are not relevant) in the continuous measure case. Should the discrete case have been represented, then each line would have been replaced with a series of steps, that would have been less comprehensive.

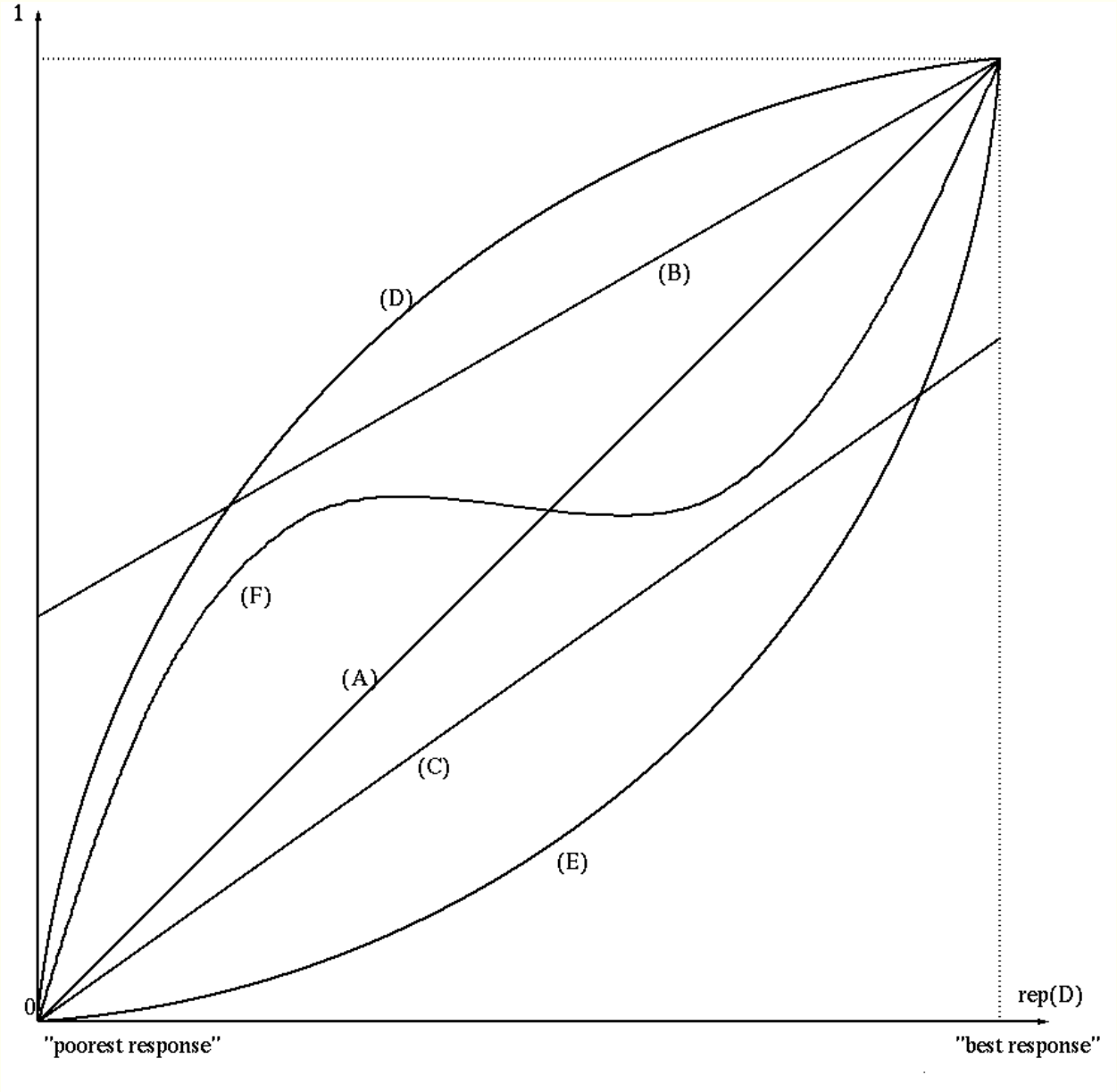

Figure 2. Shapes of various quality measures:
(a) ideal / (b) incomplete & indulgent / (c) incomplete & severe / (d) indulgent / (e) severe / (f) incoherent

Let us briefly comment the examples. The ideal measure is symbolized here by line (a). Measure (b) never attains scores close to zero, thus it does not satisfy the (LL-3) condition, and so neither (LL-2), and may be called "incomplete indulgent". In the same way, measure (c) does not satisfy the (UL) condition, as a perfect response does not obtain a 100% score. The measure that does not satisfy (UL) because it attributes 100% to imperfect responses was not represented: the situation is rare, as it is easy to detect by evaluators.

Also, measure (d) is more indulgent than measure (a), which in its turn is more indulgent than measure (e); the same relations hold for triplets (b)-(a)-(c), (d)-(f)-(e) and (b)-(f)-(c). The measures, respectively, (b)-(d), (c)-(e), (a)-(f) cannot be compared from the point of view of indulgence/severity. Finally, (f) is a non regular (non uniform) measure, which does not satisfy condition (REG); however, it could become a regular measure for a different kind of task, one that would change the appreciation of some responses.

3.4.5.5 Two applications of the framework

Information retrieval (IR)
Evaluation of IR systems [VR79][MG83] is almost exclusively concerned with their main capacity, that is, quite obviously, the capacity to retrieve the documents that are relevant to a given request. The request constitutes the input data (D_i) together with the document database (which is not likely to vary during the evaluation phase). The formal specification of this capacity cannot be formally specified; indeed, while the response format (e.g. the minimal/maximal number of documents) can be specified, there is no formal definition of "relevance" to the request – which renders the task a subject of *evaluation*.

In order to apply the present framework to the quality measures, the main capacity should be thought of as twofold: C_1 is the capacity to select documents including the relevant ones, and C_2 is the capacity to avoid non relevant documents. Evaluators agree to define as *recall* (r) the measure *mu* (C_1) defined by the average number of documents selected among all the relevant documents in the database; also, *precision* (p) is the measure *mu* (C_2) defined by the average number of relevant documents among those selected. The values of *mu* (C_1) et *mu* (C_2) for a given system are estimated (as described above) using a certain number of testing requests, for which the correct answers (i.e. the set of relevant documents in a fixed database) have been defined by experts. The *appreciation* phase is here incorporated in the *measure* phase (there is no further transformation of the scores), and the *summarization* phase is simply the computation of the harmonic mean of recall and precision, leading to a single score, called *f-measure*: $2/(1/r + 1/p)$. Summarization is imperative as there are trivial techniques that artificially increase recall *or* precision, but never both; in other words, both capacities are relevant to evaluators, so that is why an unique score is demanded.

Recall, precision and f-measure, as defined for IR, satisfy generally the coherence criteria: moreover, these measures are widely accepted by the IR community, and do not seem to have serious alternatives. Let us examine for instance the (LL) condition: for the capacity C_1, the worst responses are obviously those that do not contain any relevant document, and this is equivalent to zero recall (except in the extreme case when the response contains no documents at all). For capacity C_2, the worst responses are those containing a high proportion of non relevant documents, and this is equivalent to a very low precision (above zero if the response contains at least one relevant document).

The tactics to increase recall and precision are actually opposed: "return on average a lot of documents" vs. "return on average few documents". The *f-measure* (harmonic mean) balances these two tendencies. For instance, the trivial response which consists of *all* documents from the database obtains a non null *f-measure* (because *recall*=100% and *precision=relevant_documents/all_documents*) – but the precision is very low. The interest in choosing the harmonic mean (against, say, the arithmetic mean), is that the more one of the scores is close to zero, the more the *f-measure* is close to this score, and thus very low. The trivial response above obtains an *f-measure* close to the precision score, i.e. close to zero. Also, *f-measure* is null if and only if one of the two scores is null, i.e. no relevant document is returned.

Reference resolution
This evaluation framework has been applied, in much more detail, to (co)reference resolution [PB]. The problem of reference resolution could be expressed as grouping together a text referring expression (RE) that designate the same entity. The groups are viewed as equivalence classes for the coreference relation between REs [PB98]. Such a grouping may be expressed using SGML tags in the texts [BRU98]; so, one could verify (and even certify) for a given system

that the tags respect precise syntax. However, the fact that the tagging is correct from the point of view of designation can only be evaluated.

Evaluation consists in quantifying the resemblance between two different partitions of the RE set (for a given text), namely the key and the response. The difficulty comes from the absence of intrinsic measures for this comparison. A simplified conception of the problem uses the point of view of coreference links between REs [MUC6]. Thus, one can distinguish two capacities of the reference resolution system: find a set of links that includes the correct ones, and do not find incorrect links. A precision and a recall score could be thus defined, but several different proposals exist, yielding different results. Besides, a more refined point of view states that the comparison of sets of REs is not reducible to the comparison of links between REs.

The present evaluation framework enables precise analyses of the evaluation measures for reference resolution. For instance, most of the existing measures do not satisfy the (LL) criterion, being generally too indulgent for some special "poor responses". Some couples of existing measures have been compared through their relative indulgence. Also, the evaluation framework shows the interest of new measures for reference resolution proposed in the cited article.

3.5 Resources.

3.5.1 Evaluation Data Lifecycle.

To support evaluation on a large scale, we need to develop sound logistics for data collection, construction, annotation, storage, distribution and reuse. There are multiple problems to solve, which are rather remote from the preoccupations of technology developers, e.g. format encoding, distribution copyrights, distribution media and infrastructure, validation, etc.

The number of potential data providers and consumers is very large and keeps on increasing. Therefore, it is impractical to develop ad hoc solutions for every campaign separately. In Europe, ELRA currently plays an important role [KC98] concerning the collection, production, validation and distribution of linguistic resources. Because of its current activities and assets, ELRA is a good candidate for playing an essential role in the envisioned evaluation infrastructure.

3.5.2 Budget Estimates for Technology Evaluation.

There should be a 100% funding policy by the EC, as the evaluation campaigns are infrastructural by nature. The average cost for each of the six candidate control tasks (task 1 and 2 require much more resources than tasks 5 and 6) is estimated at 600 KEURO, corresponding to a two-year campaign. The estimates of the averages per task are:

Activity per task	*KEURO*
Production of necessary language resources (up to 4 languages)	180
Organization	90
Participants (estimated at 10 participants at 30 KEURO each)	300
Supervision	30
Total for one task	600

For the six tasks proposed, the total cost would therefore be 3.6 MEURO. These estimates were made while taking the American and three European evaluation campaigns into account.

Note that the effort devoted by DARPA to finance the Human Language Technology program is much larger than our proposal for an evaluation infrastructure in Europe. It is estimated that the funding reaches about 20 M$ per year. In average, five different tasks are conducted in parallel, both for spoken and written language processing, each costing about 4 M$ per year. But it should be stressed that DARPA fully finances the development of the systems for some participants, while in other cases DARPA restricts its financing to the organization of the campaign.

3.5.3 Budget Estimates for Usage Evaluation.

For the first campaign:

Activity per task	*EC participation*	*KEURO*
Establish the framework	100%	30
Organization	100%	90
Participants (estimated at 5 participants at 30 KEURO each)	50%	150
Supervision	100%	50
Total for one task		320

It is expected that there will be five participants in the first campaign and their costs will be 60 KEURO each. The commission will be asked to participate to these costs with a 50% contribution. The other costs will be covered by a 100% contribution.

For the following campaigns, the 30 KEURO needed to establish the framework are not necessary any more and the 50% contribution can be reduced (or even be negative if the participant pays to the campaign) depending on the success with industrial users.

Annex - Practical Considerations for Implementation

A1 The Need for a Permanent Infrastructure

Implementing the comparative evaluation paradigm in EC programs is difficult, as they are based on a call for proposals mechanism, with limited duration projects and usually a share of the cost supported by the participants. There is a need for a permanent evaluation organization of European scope, in order to cover a time scale larger than the duration of a Framework Program (FP). An ideal solution for capitalizing on the know-how gained throughout the course of several programs would be to include in the plan of action this permanent European evaluation organization, which could be responsible for defining and updating the general policy for language Technology Evaluation, for the strategic issues, for the ethical aspects, as well as for the practical organization of the evaluation campaigns (measure methodology, results computation and publication, software development, evaluation label attribution, quality control, etc.). It could either be created from scratch or by extending the mission of an existing organization. On that score, useful insights on how to take into account practical requirements like profit or not-for-profit constraints can be drawn from a parallel with successful existing organizations like the pair ELRA/ELDA. Note that ELRA already offers the means for long term capitalization on the LR produced for training and testing the systems by openly distributing them after each evaluation campaign.

A2 Selection of Evaluators and Participants

For each control task and for each evaluation campaign, there is a need for:

- communication, management, advertising and distribution;
- metrics and tools definition, implementation and application;
- gold standard definition, reference data collection or construction.

In addition to the production of specific resources, the running of an evaluation campaign will require the recruiting of both evaluators and participants. The consortium thinks that membership for both classes should be as open as possible. Evaluators ought to be selected first since they should be involved in the organization of the evaluation campaign. The proposal of any potential evaluator ought to contain at least:

- a sketch of the methodology expected to be deployed for the evaluation campaign;
- indications of the way the linguistic resources required for the campaign will be assembled in due time;
- information showing that the proponent will be able to assemble a sufficient set of participants and that he has the capability to maintain sufficient synergy throughout the whole campaign (proper communication setup, sufficient staff, recognized image as retainer of expertise in the considered domain etc.).

No restriction should be imposed on participants for participating in the dry run phase, but a selection based on the results of the dry run phase should be performed after it takes place in order to limit the number of participants in the tests to a reasonable number. This number should be fixed in advance according to the amount of resources available and advertised at the beginning of the evaluation campaign. This way of proceeding would also ensure that sufficient time is provided to solve the administrative problem that could be caused by non-EU participants, if they manage to pass the dry run test phase.

A3 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. If evaluation is deployed on large scale during FP5, the consortium advises strongly that an installation period, of 6 months at the very least, should takes place at the beginning of the program. 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). This preliminary period must be planned with care to preserve the synchrony between the evaluation campaigns and the framework programs.

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. Those 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. A first selection of 6 topics could be:

1. Broadcast news transcription;
2. Cross lingual information retrieval;
3. Text-to-speech synthesis;
4. Text summarization;
5. Language models;
6. Morphosyntactic tagging, lemmatization, word sense disambiguation.

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 according to technological 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.

A4 Evaluation in a Multilingual Context

A specific difficulty for using the evaluation paradigm in a European framework is the multilingual nature of Europe. The proposal is 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. It would be even better if all the evaluation campaigns would share a common language because the evaluation of different kinds of technologies, including complete systems and components, on the same data would then be possible. 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.

The proposal is to select a list of languages (up to 4, possibly including English) in the first step, for which there are a large enough number of potential participants, as identified by the consortia in charge of evaluation. This is in agreement with the fact that the goal is to evaluate technology, not specific applications in a given language. In a second step (future FP), other languages could be addressed, both by domestic laboratories and by those who participated in the previous evaluation campaign, and gained enough know-how in developing systems for the evaluation task to be able to easily tune their system to a new language.

A5 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 are grouped into technology clusters (either the technology used or developed in systems or components or the overall technological function implemented by the project result). Then, it is necessary to have a second call for proposals in order to select organizers for the evaluation campaign, as selecting an organizer beforehand or selecting an organizer among the projects already selected, run a very high risk of recruiting an organizer lacking the domain knowledge needed to appreciate the issues at stake, or of having an organizer with a biased opinion because of his involvement in his own project. Not mentioning the fact that organizing an evaluation is a time consuming activity which is poorly supported with the amount of resource generally allotted to project complementary issues.

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 through 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 an evaluation organizers network, 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.

References

1. [ALMPR97] G. Adda, J.Lecomte, J. Mariani, P. Paroubek, M. Rajman, Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de Parties du Discours pour le Français, Actes des 1 ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingenierie de la Langue de l'Aupelf-Uref, Avignon, Avril 1997.

2. [ALMPR98] G. Adda, J. Lecomte, J.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.

3. [ATW96] E. Atwell, Comparative Evaluation of Grammatical Annotation Models, in R. Sutcliffe, H.-D. Koch, and A. McElligott (eds), Industrial Parsing of Technical Manuals, Amsterdam, Rodopi, 1996

4. [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.

5. [BRU98] F. Bruneseaux, Noms propres, syntagmes nominaux, expressions référentielles: repérage et codage. Langues, 1(1):46-59, 1998

6. [CF98] C. Felbaum (Editor), Wordnet An Electronic Lexical Database, MIT Press, 1998.

7. [COH60] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement, 20, 37-46, 1960.

8. [COH68] J. Cohen Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, Psychological Bulletin, (70)4, 213-220.

9. [COM98] URL: http://fofoca.mitre.org/index.html

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

11. [DH98] D. Harman, The Text REtrieval Conference (TRECs) and the Cross- Language Track, in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

12. [GG97] G. Grefenstette (Editor), Cross-Language Information Retrieval, Kluwer Academic Publishers, ISBN-0-7923-8122-X.

13. [HAUS94] R. Hauser, The Coordinators' Final Report on the First Morpholympics, LDV-FORUM, vol. 11-1, June 1994, ISSN 0172-9926.

14. [HLT97] European Commission, Human Language Technology, Living and Working Together in the Information Society, Discussion Document, Luxembourg, July 1997. URL: http://www2.echo.lu/langeng/ist/hlt/paper.html

15. [HLT98] European Commission, Human Language Technology, Proposal Concerning The IST Program 1998-2002 (excerpts), COM (98) 305 Final, 13 May 1998. URL: http://www.linglink.lu/le/ist/ist/excerpts_ist_pgme.htm.

16. [ILL90] V. Illingworth, Dictionary of computing, Oxford University Press, London, 1990.

17. [IV98] N. Ide & J. Véronis, Introduction to the special issue on word sense disambiguation: the state of the art., Computational Linguistics, 24(1), 1998.

18. [AK98] A. 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.

19. [HAB93] H. Habrias, Introduction à la spécification, Masson, Paris, 1997.

20. [HAB97] H. Habrias, Dictionnaire encyclopédique du génie locigiel, Masson, Paris, 1997.

21. [ISO91] International Standard ISO/IEC 9126. Information technology -- Software product evaluation - Quality Characteristics and guidelines for their use. Geneva, International Organization for Standardization, International Electrotechnical Commission, 1991

22. [JARD98] M. Jardino, F. Bimbot, S. Igounet, K. Smaili, I. Zitouni, M. El-Bèze, A first evaluation campaign for language models, in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

23. [JD95] J. Diver, Principles and Practice of Language Engineering Project Clustering, Version 1.1 - Final, LINGLINK-LE1-1951, November 10th 1995.

24. [JM98] J. Mariani, The Aupelf-Uref Evaluation-Based Language Engineering Actions and Related Projects, First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

25. [JM99] J. Mariani & P. Paroubek, Human Languge Technologies Evaluation in the European Framework, Proceedings of the DARPA Broadcast News Workshop, February 28th-March 3rd, Herndon, Virginia, 1999

26. [JP97] J. Peckham, Bringing Language Engineering to Market, Language Engineering Concertation and Project Review, Mondorf-les-Bains, March 1998.

27. [JPSSJGTH98] J. Polifroni, S. Seneff, J. Glass, and T. Hazen, Evaluation Methodology for a Telephone-Based Conversational System, First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

28. [KC98] K. Choukri, The European Language Resource Association, in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

29. [KNRGRC95] K. Netter, R. Crouch, R. Gaizauskas, et al., Interim Report of the Study Group on Assessment and Evaluation, April 1995.

30. [KRI80] K. Krippendorff, Content Analysis: An Introduction to Its Methodology, Sage Publications, Beverly Hills, CA, 1980.

31. [KSJ95] K. Sparck Jones, J. R. Galliers, Evaluating Natural Language Processing Systems, Springer-Verlag, 1995.

32. [LAU90] J.-L. Laurière, L'intelligence artificielle : résolution de problèmes par l'Homme et la machine, Eyrolles, Paris, 1987.

33. [LG90] B. Liskov & J. Guttag, La maîtrise du développement de logiciel : abstraction et speécification, Les Éditions de l'Organisation, Paris, 1990.

34. [LGLK98] European Commission, Thematic Clustering, Language Engineering Harnessing the Power of Language, URL: http://www.linglink.lu/le/concert/clusprin.html

35. [LH98] L. Hirshman, Language Understanding Evaluations: Lessons Learned from MUC and ATIS, in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

36. [MP90] M. Minsky and S. Papert, Perceptrons - Expanded Edition, MIT Press, 1990.

37. [ML98] M. Lieberman, C. Cieri, The Creation, Distribution and Use of Linguistic Data: The Case of The Linguistic Data Consortium, in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

38. [MITR98] I. Mani, D. House, G. Klein, L. Hirschman, L. Obrsi, T. Firmin, M. Chizanowski, B. Sundheim, The TIPSTER SUMMAC Text Summarization Evaluation,

Final Report, October 1998, MTR98W0000138, MITRE Corporation. Mac Lean, Virginia, USA.
http://www.itl.nist.gov/div894/894.02/related_projects/tipster_summac/final_rpt.html

39. [MK96] M. King et al., Evaluation of Natural Language Processing Systems - EAGLES Final Report, EAG-WEG-PR.2, October 1996, ISBN-87-90708-00-8.

40. [MLIM98] E. Hovy, N. Ide, R. Frederking, J. Mariani, A. Zampolli, Editors, Multilingual Information Management – Current Levels and Future Abilities, A study commissioned by the US National Science Foundation and also delivered to the European Commission Language Engineering Office and the US Defense Advance Research Projects Agency, July 1998.
URL: http://www.cs.cmu.edu/~ref/mlim/

41. [MS58] J.-G. March & H.-A. Simon, Organizations, John Wiley, New York, 1958.

42. [MUC6] Proceedings of the 6th Message Understanding Conference (DARPA MUC-6 '95), Morgan Kaufman, San Fransico, CA, 1995

43. [PB] A. Pobescu-Belis, Évaluation numérique de la résolution de référence : critiques et propositions, Traitement Automatique des Langues (TAL), in print.

44. [PB98] A. Pobescu-Belis, I. Robba and G. Sabah, Reference Resolution Beyond Coreference: a Conceptual Frame and its Application. In COLING-ACL '98, Montréal, Québec, Canada, Université de Montréal, 1998.

45. [RL98] R. Lockwood, Language Technology: Understanding the Market, Language Engineering Concertation and Project Review, Mondorf-les-Bains, March 1998.

46. [RS97] R. Sproat (Editor), Multilingual Text-To-Speech Synthesis - The Bell Labs Approach, Kluwer Academic Publishers, ISBN 0-7923-8027-4.

47. [RY97] P. Resnik and D. Yarowsky, A perspective on word sense disambiguation methods and their evaluation, position paper presented at the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, held in conjunction with ANLP-97 in Washington, D.C., USA, April 4-5, 1997.

48. [SMG83] G. Salton & M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

49. [SOM92] I. Sommerville, Le génie logiciel, Addison-Wesley, Paris, 1992.

50. [SYLC98] S. Young, L. Chase, Speech Recognition Evaluation, A Review of the US CSR and LVCSR Programs, Journal of Computer Speech and Language, 1998.

51. [SYLL97] S. J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.-L. Gauvain, D.J. Kershaw, L. Lamel, D. Leeuwen, D. Pye, A.J. Robinson, H.J.M. Steeneken, an P.C. Woodland, Multilingual Large Vocabulary Speech Recognition: The European SQALE Project, Computer Speech and Language, Vol. 11, 1997.

52. [VR79] C. J. Van Rijsbergen, Information Retrieval, Butterworth, London, 1979.

53. [VER98] J. Véronis, A Study of Polysemy Judgement and Inter-Annotator Agreement, Advanced paper, presented at the Pilot SENSEVAL workshop in Herstmonceux, UK, 2-4 September, 1998.

54. [WGIMNSYZ98] S. Wegman, L. Gillick, Y. Ito, L. Manganaro, M. Newman, F. Scattone, J. Yamron, P. Zhan, Dragon System Automatic Transcription System for the New TDT Corpus, in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

55. [WLKA97] M. Walker, D. Litman, C. Kamm, A. Abella, PARADISE: A Framework for Evaluating Spoken Dialogue Agents, In Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, ACL 97, 1997.

56. [WHFFM97] M. Walker, D. Hindle, J. Fromer, G. Di Fabbrizio, G. Mestel, Evaluating Competing Agent Strategies For A Voice Email Agent, in Proceedings of the 5th European Conference On Speech Communication And Technology, Rhodes, September, 1997.