Principles of Evaluation in Natural Language Processing

Patrick Paroubek[1] — **Stéphane Chaudiron**[2] — **Lynette Hirschman**[3]

LIMSI - CNRS, Bât. 508 Université Paris XI
BP 133 - 91403 ORSAY Cedex - France pap@limsi.fr [1]

GERiiCO, Université Charles-de-Gaulle (Lille 3)
B.P. 60149, 59 653 Villeneuve d'Ascq Cedex, France
stephane.chaudiron@univ-lille3.fr [2]

The MITRE Corporation, 202 Burlington Rd., Bedford, MA, USA
lynette@mitre.org [3]

ABSTRACT. *In this special issue of TAL, we look at the fundamental principles underlying evaluation in natural language processing. We adopt a global point of view that goes beyond the horizon of a single evaluation campaign or a particular protocol. After a brief review of history and terminology, we will address the topic of a gold standard for natural language processing, of annotation quality, of the amount of data, of the difference between technology evaluation and usage evaluation, of dialog systems, and of standards, before concluding with a short discussion of the articles in this special issue and some prospective remarks.*

RÉSUMÉ. *Dans ce numéro spécial de TAL nous nous intéressons aux principes fondamentaux qui sous-tendent l'évaluation pour le traitement automatique du langage naturel, que nous abordons de manière globale, c'est à dire au delà de l'horizon d'une seule campagne d'évaluation ou d'un protocole particulier. Après un rappel historique et terminologique, nous aborderons le sujet de la référence pour le traitement du langage naturel, de la qualité des annotations, de la quantité des données, des différence entre évaluation de technologie et évaluation d'usage, de l'évaluation des systèmes de dialogue, des standards avant de conclure sur une bref présentation des articles du numéro et quelques remarques prospectives.*

KEYWORDS: *evaluation, gold standard, language technology, usage, dialog system*

MOTS-CLÉS : *évaluation, référence, technologie du langage, usage, système de dialogue*

1. Introduction

1.1. *A bit of history*

For a long time talking about evaluation was a forbidden topic (King, 1984) in the natural language processing (NLP) community because of the ALPAC (S. Nirenburg and Wilks, 2003) report which had generated a long and drastic cut in funding for research in machine translation in the United States. The first sign of a possible change of mind came in 1987, again from America, with the organization of a series of evaluation campaigns for speech processing (Pallett, 2003), then for text understanding –for a survey of evaluation in the domain see TIPSTER[1] (Harman, 1992) program. A few years later, TREC[2] (Voorhees and Harman, 2005) was born to address the needs of the information and document retrieval research community. It was the first of an ongoing series of evaluation campaigns on information retrieval that continues until today. Afterwards, the importance of evaluation for the field kept growing, along with the number of campaigns, the number of participants and the variety of tasks, until one could speak of the "evaluation paradigm" (Adda *et al.*, 1998).

People in Europe were more hesitant about evaluation campaigns, since to our knowledge the first event of the sort happened in 1994 in Germany with the "morpholympics" (Hauser, 1994) on morphological analyzers for German. The same year the GRACE (Adda *et al.*, 1998) campaign on Part-Of-Speech taggers of French was started in France. Among the reasons we can put forward for this late and more tentative rebirth of evaluation in Europe there are : the nature of the funding agencies, the economic and geopolitical contexts and the possibility for Europeans to participate in American campaigns. Nevertheless, evaluation regained little by little some status also in Europe as attested by the 7 campaigns of the FRANCIL program (Chibout *et al.*, 2000) for text and speech, the series of self-supported campaigns Senseval on lexical semantics organized by the ACL-SIGLEX working group (Edmonds and Kilgarriff, 2003), its follow-up Semeval (Agirre *et al.*, 2007) or the more recent evaluations campaigns for Portuguese text analysis (Santos *et al.*, 2003) (Santos and Cardoso, 2006), as well as examples of national programs on evaluation like TECHNOLANGUE [3] (Mapelli *et al.*, 2004) in France with the 8 evaluation campaigns on both speech and text of the EVALDA project or the latest EVALITA (Magnini and Cappelli, 2007) in Italy with its 5 campaigns on text analysis. The picture is even more encouraging if you look at European project which have addressed the subject of evaluation within the past few years, from EAGLES (King *et al.*, 1996) to the CLEF evaluation series (Agosti *et al.*, 2007). In figure 1 some of the salient evaluation related events mentioned in this article are located on the time line.

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster

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

[3] http://www.technolangue.net

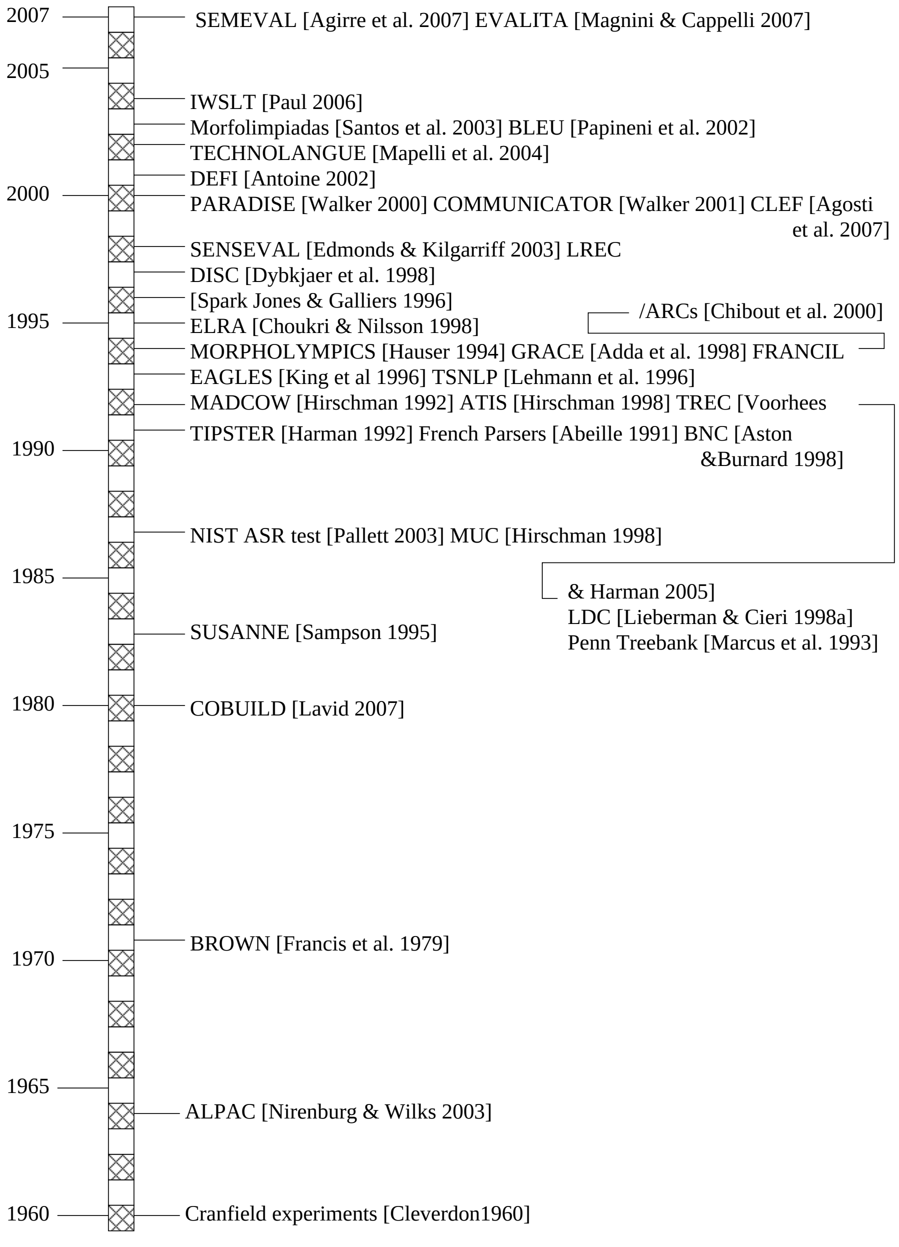
2007
SEMEVAL [Agirre et al. 2007] EVALITA [Magnini & Cappelli 2007]
2005
IWSLT [Paul 2006]
Morfolimpiadas [Santos et al. 2003] BLEU [Papineni et al. 2002]
TECHNOLANGUE [Mapelli et al. 2004]
DEFI [Antoine 2002]
2000
PARADISE [Walker 2000] COMMUNICATOR [Walker 2001] CLEF [Agosti et al. 2007]
SENSEVAL [Edmonds & Kilgarriff 2003] LREC
DISC [Dybkjaer et al. 1998]
[Spark Jones & Galliers 1996]
1995
ELRA [Choukri & Nilsson 1998]
/ARCs [Chibout et al. 2000]
MORPHOLYMPICS [Hauser 1994] GRACE [Adda et al. 1998] FRANCIL
EAGLES [King et al 1996] TSNLP [Lehmann et al. 1996]
MADCOW [Hirschman 1992] ATIS [Hirschman 1998] TREC [Voorhees
TIPSTER [Harman 1992] French Parsers [Abeille 1991] BNC [Aston &Burnard 1998]
1990
NIST ASR test [Pallett 2003] MUC [Hirschman 1998]
1985
& Harman 2005]
LDC [Lieberman & Cieri 1998a]
Penn Treebank [Marcus et al. 1993]
SUSANNE [Sampson 1995]
1980
COBUILD [Lavid 2007]
1975
BROWN [Francis et al. 1979]
1970
1965
ALPAC [Nirenburg & Wilks 2003]
1960
Cranfield experiments [Cleverdon1960]

Figure 1. *Salient events related to evaluation mentioned in this article (for evaluation campaign series, e.g. like TREC, only the first event is mentioned).*

1.2. *Some Evaluation Terminology*

In NLP, identifying in a complete system a set of independent variables representative of the observed function is often hard, since the functions involved are tightly coupled. When evaluating, the need to take into account the operational setup adds an extra factor of complexity. This is why (Sparck Jones and Galliers, 1996), in their analysis and review of NLP system evaluation, stress the importance of distinguishing evaluation criteria relating to the language processing objective (*intrinsic* criteria), from the ones relating to its role with respect to the purpose of the whole setup (*extrinsic* criteria). One of the key questions is whether the operational setup requires the help of a human, in which case evaluation will also have to take into account human variability in the test conditions (Sparck Jones, 2001). The European project EAGLES (King *et al.*, 1996) used the role of the human operator as a guide to recast the question of evaluation in terms of users' perspective. The resulting evaluation methodology is centered on the consumer report paradigm and distinguishes three kinds of evaluation:

1) *progress evaluation*, where the current state of a system is assessed against a desired target state,

2) *adequacy evaluation*, where the adequacy of a system for some intended use is assessed,

3) *diagnostic evaluation*, where the assessment of the system is used to find where it fails and why.

Among the other general characterizations of evaluation encountered in the literature, the following ones emerge as main characteristics of evaluation methodologies (Paroubek, 2007):

1) *black box* evaluation (Palmer and Finin, 1990), when only the global function performed between the input and output of a systems is accessible to observation,

2) and *white box* (Palmer and Finin, 1990) evaluation when sub-functions of the system are also accessible,

3) *objective* evaluation, if measurements are performed directly on data produced by the process under test,

4) *subjective* evaluation if the measurements are based on the perception that human beings have of such process,

5) *qualitative* evaluation when the result is a label descriptive of the behavior of a system,

6) *quantitative* when the result is the value of the measurement of a particular variable,

7) *technology* when one measures the performance of a system on a generic task (the specific aspects of any application, environment, culture and language being abstracted as much as possible from the task),

8) *user-oriented* evaluation, another trend of the evaluation process which refers to the way real users use NLP systems while the previous trends may be considered as more "system oriented". Nevertheless, this distinction between system and user oriented is not so clear and needs to be clarified, which is the purpose of sections 4 and 5.

Data produced by the systems participating in an evaluation campaign are often qualified as "hypothesis" while data created to represent the *gold-standard* (Mitkov, 2005) are labeled "reference".

According to now-acknowledged quality criteria, an evaluation campaign should comprise four phases:

1) The training phase: distribution of the training data so the participants can calibrate their system to the test conditions.

2) The dry-run phase: first real-life test of the evaluation protocol with a (generally small sized) gold-standard data set. Although, they are communicated to the participants, the performance results are not considered as valid, since the dry-run may have revealed things that need to be adjusted in the protocol or in the participants' systems.

3) The running of the actual evaluation with the full gold-standard data set to compute the performance results.

4) The adjudication phase: validation by the participants of the results produced in the test phase. In general, this phase ends with the organization of a (possibly private) workshop where all the participants present their methods and their systems and discuss the results of the evaluation.

2. Language and the multiplicity of gold standards

Is it possible to agree on a common reference when language is concerned? This issue is more salient when the evaluation metrics depend directly on the ability of the system to emulate text understanding or text generation – for instance in information extraction, automatic summarization or machine translation, as opposed to tasks where the metrics are indirectly dependent on these abilities as is the case for annotation tasks, e.g. Part Of Speech tagging. Given a text to translate from one language to another, it is impossible to propose a particular translation as a gold standard since there are so many different ways to phrase a meaning. Even if we could come up with a set of universal quality criteria for evaluating a translation, we would still be far from the mark since we would still lack the interpretative power to automatically apply those criteria to define a unique gold standard; up to now the best that was achieved in that direction for machine translation was BLEU (Papineni *et al.*, 2002), an evaluation metric that computes a text distance based on trigrams and shows correlate with human evaluation results, but the controversy about it is lively. For annotation tasks, it is much more easier to come up with a unique annotation of a given text inside a particular theoretical framework; there even exist quality tests like the Kappa coefficient which measures a distance between the observed agreement and the agreement expected to happen by chance. In the case of annotation tasks, the challenge is for a community to agree on a unique theoretical framework, as opposed to coping with language variability. For instance, the decision about whether to annotate a past participle as a verbal form or as an adjectival form or as belonging to a category that pertains to both classes depends on the underlying theoretical framework, but the recognition of the past participle can be accomplished by machines with the level of human performance.

Also in relation with the multiplicity of gold standards, there is question of whether the performance of a language processing system should be measured against a theoretical objective (the maximal performance value defined by the evaluation metrics), or rather against the average performance level displayed by humans when performing the task under consideration, as (Paek, 2001) proposes to do when evaluating spoken language dialog systems.

3. On the quantity and quality of annotation

In all the domains of NLP, the evaluation practices evolve according to the same pattern. At first, evaluation is done by human experts who examine the output or behavior of a system when it processes a set of test sentences. A good historical example of this kind of practice is offered by the first comparative evaluation of parsers of French (Abeillé, 1991) or the first competition of morphological analyzers for German, the Morpholympics (Hauser, 1994). For particular domains like speech synthesis, this is almost the only way to consider evaluation; also simpler evaluation protocols based on text to phoneme transcription have been used in the past. Very often this way of performing evaluation implies the use of an analysis grid (Blache and Morin, 2003) which lists evaluation features. For instance DISC (Dybkjær *et al.*, 1998) was a European project which produced such feature set for spoken language dialog systems. Such an evaluation protocol requires no reference data, since the only data needed are input data.

But to limit the bias introduced by a particular group of experts and to promote reuse of linguistic knowledge, one often creates test suites which objectify the experts' knowledge and can be considered as the next evolutionary step in the development of the evaluation paradigm in a particular domain. For instance in the case of parsing, the European project TSNLP (Lehmann *et al.*, 1996)(Oepen *et al.*, 1996) was built for a set of European languages to contain both positive and negative parsing examples, classified according to linguistic phenomena involved. As opposed to the straightforward expert examination, which does not require any data apart from the input one, test suites require a relatively small amount of output data but with very high quality annotations since their aim is to synthesize expert knowledge about a given processing of language. Although they are of a great help to experts and developers, test suites do not reflect the statistical distribution of the phenomena encountered in real corpora and they are also too small to be reused for evaluation (except for non-regression tests), because once they have been disclosed, it is relatively easy to customize a system for the specific examples contained in the test suite.

It is at this moment that often corpus based evaluation enters the picture, where the field has matured enough to have available a relatively large amount of annotated data for comparative evaluation or where the data is created especially for evaluation purposes, a practice that led to the creation of the Linguistic Data Consortium (Liberman and Cieri, 1998a). The most famous corpora are certainly the Brown corpus (Francis *et al.*, 1979), the SUSANNE corpus (Sampson, 1995), COBUILD (Lavid, 2007), the BNC (Aston and Burnard, 1998) and the Penn Treebank (Marcus *et al.*, 1993), which have inspired many other developments like (Brant *et al.*, 2002), or (Abeillé *et al.*, 2000) for French. But corpus based approaches are far from solving all the problems since they constrain the system developers to use the annotation formalism of the evaluation corpus, and they are not adapted to interactive systems evaluation. We will address both issues respectively in sections 6 and 7. Furthermore, if corpus based evaluation methods are an answer to the distributional representation problem since they offer a large enough language sample, they suffer from a correlated weakness: how to ensure consistency of the annotations throughout the whole corpus? The question of the the balance between the amount of data annotated against the quality of the annotation can be separated into the following three questions:

1) What is the amount of data required to capture a sufficient number of the linguistic events targeted by the evaluation at hand in order to be able to produce relevant performance measures?

2) What is the minimal quality level needed for the evaluation corpus to produce relevant performance measures?

3) how to achieve consistent annotation of a large amount of data at low cost?

The first question is an open question in NLP for all corpus based methods, and despite the arguments provided by some that the more data the better (Banko and Brill, 2001), the only element of proof forwarded so far have concerned very basic language processing tasks.

The second question raises the question of the utility of the evaluation itself. Here again, this is an open question since a reference corpus may be of a quality level insufficient to provide adequate learning material while at the same time being able to produce useful insights to system developers when used in an evaluation campaign.

Finding a solution to the third question is equivalent to finding a solution for the task which is the object of the evaluation if we look for a fully automatic solution. And of course, the evaluation tasks are precisely chosen because they pose problems.

4. Technology oriented evaluation

Technology is defined in the TLFI[4] (Pierrel, 2003) as *the systematic study of processes, methods, instruments or tools of a domain or the comparative study of techniques*, while in a the Meriam-Webster Online [5] it is *the practical application of knowledge especially in a particular area: engineering*. Where the French definition uses terms like "systematic study" or "comparative study", the English one mentions "engineering", a field where the notions of measure, benchmarking and standards are prominent. We can see, in the use of methods yielding synthetic results that are easy to grasp by non-experts, one of the reasons behind the success (Cole *et al.*, 1996) of the re-introduction in NLP of technology oriented evaluation by NIST and DARPA. In their recurrent evaluation campaigns, language applications were considered as a kind of technological device and submitted to an evaluation protocol which focused on a limited number of objective quantitative performance measures. In addition to measure, the qualifier "technology" means also standards and reusability in different contexts, thus the term "component technology" used sometimes (Wayne, 1991), e.g. speech transcription, which is one of the components of any spoken language dialog systems (see figure 2).

In essence, technology evaluation uses *intrinsic* (Sparck Jones and Galliers, 1996) evaluation criteria, since the aim is to correlate the observed performance with internal parameter settings, remaining as much as possible independent of the context of use. But more than the simple ability to produce a picture of a technological component at a particular time, it is the repetition of evaluation campaigns at regular intervals on the same topics using similar control tasks (Braschler and Peters, 2003) that led to the success of deployment of technology evaluation in the US, because it provided clear evidence that the funding spent had a real impact on the field by plotting performance curves showing improvement over the years, e.g. the now famous downslope curves of automatic speech transcription error rates (Wayne, 1991).

A second reason for the success of the US evaluations was the openness of the campaigns; for most of them there was no restriction attached to the participation apart from having an operational system and adhering to the rules set for the campaign. Although technology evaluation is now widely accepted in NLP as attested by the growing number of evaluation campaigns proposed every year to systems developers abroad, no permanent infrastructure (Mariani and Paroubek, 1999) has yet been deployed elsewhere than in the US (Mariani, 2005). Periodic programs have occurred, e.g., in France with TECHNOLANGUE, in Italy with EVALITA (Magnini and Cappelli, 2007), or in Japan (Paul, 2006), but Europe is still lacking a permanent infrastructure for evaluation.

[4] see http://atilf.atilf.fr/tlf.htm, *«Science des techniques, étude systématique des procédés, des méthodes, des instruments ou des outils propres à un ou plusieurs domaine(s) technique(s), art(s) ou métier(s). La technologie, ou étude comparative des techniques,»*

[5] http://www.merriam-webster.com/

5. User oriented evaluation

The use of the term “user-oriented” is quite problematic by itself because of its polysemy according to the different scientific communities. The role and the involvement of real users in evaluation campaigns may differ quite deeply. In a certain usage, “user-oriented” may be just defined as the attention given to users’ behavior in order to integrate some individual or social characteristics in the evaluation protocol and to be closer to the “ground truth”. For example, in a information filtering campaign, technological trackers may be asked to design the profiles to be used by the systems instead of having the profiles created by non practitioners. In a machine translation campaign, real translators may be asked to give relevance judgments to the texts translated. More generally, as shown in these examples, users participate in the evaluation process as experts for a domain and their role consists of improving the protocol to be closer to the “ground truth”. In this approach, evaluation is still system oriented but it tries, to some extent, to take into account the context of use and some behavioral characteristics of the users.

Another way to define what can be a “user-oriented” evaluation process is to consider a new paradigm where the goal is not to improve the performance of the systems but to analyze how users utilize NLP software in their environment, how they manage the various functionalities of the software, and how they integrate the software in a more complex device. Therefore, the goal is to collect information on the usage of NLP systems, independently of the performance of the systems. Following D. Ellis’ statement (Ellis, 1992) concerning the Information Retrieval (IR) communities, two major paradigms may be identified for NLP evaluation: the physical (system oriented) and the cognitive[6] (user oriented) one. Most researchers and evaluation specialists would agree on this basic distinction even if the term “user-oriented” needs to be defined more closely. Early work in NLP emphasized the technical part of the linguistic process by concentrating in particular on improving the algorithms and the coding schemes for representing the text or the speech to be automated. Even now, performance continues to be measured in terms of a systems’ ability to process a document and many protocols still use the precision and recall ratios. Coming from the information retrieval (IR) evaluation effort in the earlier days with the Cranfield experiments (Cleverdon, 1960), these measures are widely used in spite of the numerous theoretical and methodological problems that some authors pointed out (Ellis, 1990) (Schamber, 1994). This focus continues to the present with its most visible manifestation the series of TRECs (Voorhees and Harman, 2005).

Given the limitations of the system oriented paradigm, a new approach could be identified by the late eighties, with a specific interest in users and their behaviors. Two separate directions can be identified: one was originally an attempt to incorporate the user more explicitly within the system paradigm with the goal of improving the performance of the NLP systems, and the other stressed on the user as a focus in

[6] We will not discuss here the fact that the way Ellis defines the term "cognitive" is much wider than the ordinary acceptance in cognitive science.

itself. This shift came partially as a result of considering anew some of the underlying theoretical aspects of the system paradigm, i.e., the representation of the linguistics resources (grammars, dictionaries), the design of the systems, and the components of the processing. It came also from the reconsideration of the role of the user in the acceptance of the systems and the fact that different users might have different perceptions of the quality of the results given by the systems, the efficiency of the systems, the relevance of the processing, and the ability of the systems to match with the real users' needs.

A strong impetus for this shift was the belief that, if it is possible to understand the variables that affect a user's performance with a given system, it would be easier to design systems that worked better for a wide variety of users by taking into account their individual characteristics. Roughly, three main directions may be pointed out. A first group of researchers are specifically interested in understanding some central concepts used in the evaluation approaches, such as quality, efficiency, and in particular, the concept of relevance and the relevance judging process which are considered as key issues in evaluating NLP systems. A second group employs cognitive science frameworks and methods to investigate individual characteristics of users which might affect their performance with NLP systems: user behavior and acceptability of the systems. A third group investigates the use of NLP systems as a communication process and employs qualitative methods derived from sociology, ethnomethodology, and anthropology.

The concept of relevance is very central in the IR process (see in particular (Saracevic, 2007) but is now widely discussed for extraction tools, machine translation and so on. The nature of relevance and how to judge it has been a key question in IR evaluation since the first evaluation campaigns in the early sixties (the Cranfield tests). From the need to determine the relevance of documents to queries, a full discussion of the variety of methods employed for achieving more consistent relevance judgments has develops and still continues. (Schamber, 1994) and (Saracevic, 2007) have summarized much of the discussion for the IR community but we also find in (Sperber and Wilson, 1989) a more philosophical viewpoint on the question.

Much of the early work in relevance judgments investigated the conditions under which judgments changed, in order to determine better methods for generating the set of relevant documents to be used for computing precision and recall. Even today, evaluation campaigns such as the INFILE [7] campaign discusses the best way to integrate users considerations in the protocol. The user oriented researchers also focused on the extensive literature on changing relevance and have attempted to express why and how these judgments change. These works have led to a widely shared understanding that relevance judgments change over time, over the different contexts of use, and for different categories of users according to socio-professional situations and individ-

[7] Started in 2007, INformation, Filtrage, Evaluation is a cross-language adaptive filtering evaluation campaign, sponsored by the French National Research Agency which extends the last filtering track of TREC 2002.

ual characteristics. These works on relevance and relevance judging provide valuable insights into NLP systems users' behavior as dynamic and situated in a particular moment in time-space. The practical implications of this reconceptualization of relevance have been discussed. For example, (Harter, 1996) tried to formalize how to take into account the changing relevance judgments in appropriate evaluation measures. But, despite these attempts, it should be noted that many efforts are still to be done in order to integrate real users behavior within the various evaluation protocols.

Another trend focusing on "user-centered" evaluation concerns works which utilize cognitive theories, frameworks and methods of perceptual processes, higher-level cognitive processes and individual differences. Generally, researchers using these approaches specify characteristics of the user (such as cognitive style) which are measured prior to an interaction with a NLP system and which are assumed to remain constant throughout the interaction. After the interaction, user performance is assessed by measures such as error rate, time elapsed, or number of texts processed. In the cognitive psychology studies, focus is placed on different two kinds of variables. First, there are the independent variables such as the ones related to the user (cognitive or learning style, experience, gender, intelligence, knowledge, personality, experience?); the ones related to the system (interface type, highlighting style, labeling style, type of display, window size); and the ones related to information (text length, mono- or multi-linguality, type of discourse, type of information).

The second type of variables are the dependent variables such as the accuracy of the process computed (error rate, readability), the process (number of commands used, number of screens accessed, time elapsed, learning curve), global measures (attitude, perception of ease of use, usefulness, perseverance, satisfaction). A major interest of researchers in this field concerns knowledge and cognitive models (mental models, world, representation of the system, ability to perform a task, domain knowledge), cognitive processes (cognitive load, cognitive behaviors, learning problem solving, memory, cognitive abilities and cognitive styles). These studies highlight the complexity of the tasks performed. Each study, using a different theoretical perspective, found some significant relationship between the independent and dependent variables. None of them, however, offers a complete and stable framework of users' behaviors to be integrated in a evaluation campaign. For example, a limitation comes from the model of the cognitive abilities which assume a clear distinction between expert, semi-expert and novice. Even if this categorization of users is easy to use and has led to numerous experimentations, some doubts arise when generalizing the results.

The third trend refers to different approaches which consider NLP systems as communication devices and focus on the user's movement through the situation. Research within this direction demonstrates the use of a very diverse set of theories and methods, making it difficult to summarize the approach succinctly. Quoting (Mey, 1977), (Ellis, 1992) pointed out the key principle driving these works: *"that any processing of information, whether perceptual or symbolic, is mediated by a system of categories or concepts which, for the purposes of the information processing device, are a model of the world"*. This assumption, which can be verified for all automatic devices, is of particular importance for the NLP systems which are based on linguistics resources designed according to a particular vision of the world. There is no need to repeat, for example, the debates concerning terminologies, ontologies or semantic networks (such as WordNet); linguistic tools are based and designed on particular visions of the world, deeply influenced by cultural, religious or ideological heritage. More generally, the studies within this "communication approach" focus on some key issues as the symbolic aspects of users' representation of technologies, people's ability or inability to communicate through a computational device, users interaction, the usage and the non-usage of NLP software. The communication studies conducted empirical works and provided several theoretical frameworks and models of the user behavior. Even if most of the models were developed to better understand the IR process, many of them address a much wider scope and try to model the whole communication process. A recent survey of user information behavior papers identified more than 70 different models or frameworks (Fisher *et al.*, 2005). One of the key questions concerning these models is their relevance and their operational effectiveness in understanding user complexity in given situations. While these works may be useful to the user modeling approach in the long-term, the limited utility of current modeling efforts for the purpose of evaluation has been pointed out in (Sparck Jones, 1990). In summary, user modeling continues to be a focus for researchers and is interesting for improving software performance particularly when systems are used by practitioners or restricted communities which are easier to model. In addition to the user modeling approach, there are a wide variety of other frameworks suggested or employed by other researchers.

A last framework presented here comes from a particular sub-domain of sociology which focuses on usage and specifically on the use of Information and Communication Technologies (ICT). Given the limitations of the early ICT systems by the beginning of the eighties, researchers were asked to identify bottlenecks, barriers and limits of the domain. A new field emerged employing qualitative methods to better understand the acceptance or the refusal of ICT systems by users. This approach is still quite prolific and gives a very interesting framework for "user-centered" evaluation, even if it does not fit easily in the evaluation metric frame. For an overview of these works, see (Jouet, 2000). In this brief section on "user oriented" studies, we first tried to point out the origins and the complexity of the term "user oriented" and the wide variety of approaches which refer to it. "User oriented" evaluation is therefore a much more complex domain that it seems to be on the developer side. Even if these works seem to be too far away from a quantitative approach of evaluation using well established protocols and metrics, we think that the two paradigms could be much closer in order to really integrate users in the evaluation protocols.

6. Dialog system evaluation

Spoken language dialog systems (SLDS) stand apart from other systems since they incorporate by necessity almost all possible kinds of basic natural language processing functionality in a single system, from both text and speech domains as the figure 6 (Lamel *et al.*, 2000) shows. Note that each functionality need not be implemented as a module of its own, but the processing path from input to output has to be complete for the system to be operational. More detailed information is available on page 117 of this issue in the article of J. Allemandou et al. that presents SIMDIAL. The number of sub-domains of natural language processing involved in the realization of a SLDS poses one of the greatest challenges of evaluation, since one is faced with the problem of correlating global evaluation measurements characterizing the relationship between input and output of the SLDS with characterizations of the working of each individual functionality, in relation with the way it is implemented in the actual system. For instance, the implementation of the dialog management functionality may be spread over several modules, making the tracking of its functioning quite difficult, even in the case of white box evaluation where evaluation is a priori made easier by the fact that one has access to the individual module inputs and outputs. Concerning individual

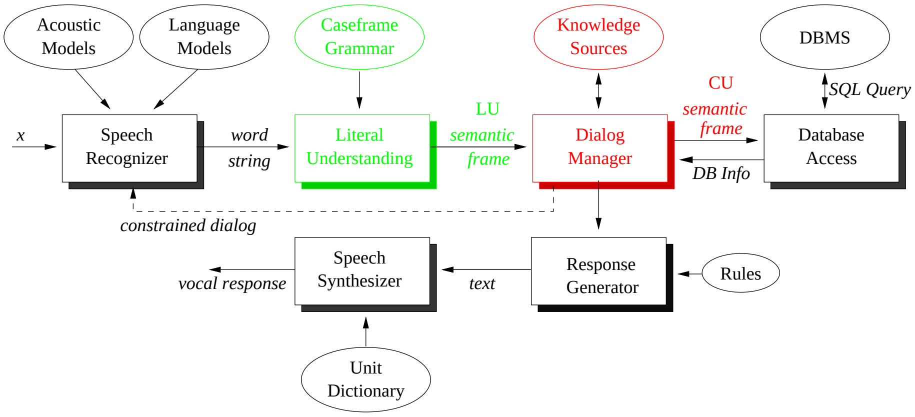

Figure 2. *End to end generic functional architecture of a spoken language dialog system.*

functionalities of SLDS, the historical development of evaluation protocols follows roughly the progression of information in a SLDS from input to output, the accent being first put on speech recognition (Burger *et al.*, 1998) and later on speech synthesis. On may offer two explanations for this fact; the first one is that speech recognition is the first module in the signal processing path, so logically the it received attention first. The second is that the more one goes toward the end of the function input-output path in a SLDS, the more it is difficult to abstract from subjective perceptual human factor in the evaluation process. For instance, the Word Error Rate measure provides a sufficient evaluation criteria for gauging speech processing technology at a level of quality which makes possible the building of an operational SLDS (e.g., train ticket reservation) and Word Error Rate does not account directly for the perceptual characteristics of the input speech signal such as intonation and prosody. But evaluation of speech synthesis needs to take into account more perceptual dimensions since its evaluation, for instance, cannot be based only on intelligibility of the information generated, since the human quality of synthesized speech is essential in the acceptability of a speech synthesizer by humans. As a consequence, we think that speech recognition technology benefited earlier from the accelerating factor brought by comparative technology evaluation, while speech synthesis took longer to get the benefits, since evaluation procedures had to have enough time to mature in order to take into account human perceptual aspects.

But for the evaluation of spoken dialog systems, there are presently no common standard methodologies or practices agreed upon by the scientific community, since the dynamic and interactive nature of dialog makes it difficult to construct a reference corpus of dialogs against which systems may be evaluated. The first large efforts toward providing an end to end evaluation protocol for an SLDS came from the USA with the ATIS (Hirschman, 1998) evaluation campaign of the MADCOW (Hirschman *et al.*, 1992) program which addressed the evaluation of an SLDS for air travel logistics database management. Later, the initiative came again from the same part of the world with the innovative protocol PARADISE (Walker *et al.*, 2000) deployed in the COMMUNICATOR project (Walker *et al.*, 2001b) which this time was testing a multimodal real-time planning and travel information management system (Walker *et al.*, 2001a). PARADISE attracted a lot of notice since it went relatively far in providing a task independent evaluation protocol, by correlating user satisfaction with objectively measurable performance parameters (Dybkjæer *et al.*, 2004).

Proposals were also made in Europe, at different scales and without the same amount of support as the US programs. Various influential projects have tried to build the foundations of an evaluation methodology for spoken dialog systems, including the European EAGLES projects (Dybkjaer 1998) and then later DISC (Giachim 1997) and SUNDIAL (Gibbon 1997); in France specific evaluation campaigns also addressed SLDS evaluation: the French speaking project AUF-Arc B2 (Mariani 1998) and the evaluation carried out by DEFI (Antoine *et al.*, 2002).

In the French TECHNOLANGUE program, the MEDIA campaign used PEACE, an evaluation protocol (*Paradigme d'Evaluation Automatique de laomprEhension hors- et en- contexte dialogique*) (Devillers 2002, Maynard 2000) crafted by the MEDIA project. This protocol proposes to separate the context-dependent and independent understanding capability evaluations of a dialog system, while using an automatic comparative diagnostic evaluation methodology. It is based on the construction of reproducible test suites from real dialogs. This paradigm takes on the idea of the DQR (Antoine 2000) and DEFI (Antoine 2002) evaluations based on test suites. The evaluation environment relies on the premise that, for database query systems, it is possible to construct a common semantic representation to which each system is capable of converting its own internal representation. Classically, context independent evaluation is carried out by comparing the interpretation produced by the systems to a reference interpretation for a given set of independent utterances. For context dependent evaluation, the context of each test utterance is artificially simulated by paraphrasing and the systems have to take it into account before proposing an interpretation for the utterance. Then their interpretation is compared with the reference interpretation. The originality of the MEDIA contribution was in the use of a paraphrased context and it benefited the French community of spoken language dialog system developers by making available both a common reference annotation scheme and the associated corpora for a tourist information task (Bonneau-Maynard *et al.*, 2006). Since that time, work on SLDSs evaluation has continued on in France and the latest achievements are presented in the article of J. Allemandou, L. Charnay, L. Devillers, M. Lauvergne and J. Mariani, page 115, with SIMDIAL, an evaluation paradigm to evaluate automatically an SLDS by means of a deterministic simulation of users.

7. Standards and evaluation

Since evaluation aims at providing a common ground to compare systems and approaches, it is by its nature an activity that is both a source and a user of standards. It is important to note that depending on the language or the professional communities you look at, the two notions of "standard" and "norm" may have different meanings. In what follows, we will use the Webster Online dictionary definitions and call a standard "*something established by authority, custom, or general consent as a model or example*" and a norm "*an authoritative standard*". A standard emerges as a fact of a community; as such it can be seen as the solution of a consensus problem, whose formal study started in the 60's (Olfati-Saber *et al.*, 2007) and which has even been proposed as a possible model of language emergence among robots (Kaplan, 2001). But we are more interested here in the relationship that evaluation entertains with standards than in the modeling of the standardization process itself. So for what concerns us, standards deal with three aspects of evaluation : the evaluation protocol, the evaluation metrics and the annotation of corpora.

One cannot deny that the renewal of evaluation campaigns initiated in the USA for speech processing (Pallett, 2003), text understanding (Grishman and Sundheim, 1996) and information retrieval (Voorhees and Harman, 2005), supported by a long standing effort over many years, helped to establish some evaluation procedures used in these campaigns as standards. For instance, (Voorhees, 2002) explains why and how test collections became a standard practice for evaluating information retrieval systems. The American campaign left their imprint on many evaluation campaigns, so much that most of the evaluation campaigns addressing technology evaluation that take place nowadays follow the four steps plan described in page 10 of this article.

Also popularized by the TREC campaigns are the Precision and Recall measures, which now considered standard for Information Retrieval (Manning and Schütze, 2002), or the Mean Reciprocal Rank measure for evaluating Question Answering systems (Jurafsky and Martin, 2000). Word Error Rate and Perplexity (Chen *et al.*, 1998) are two measures that became standards in the speech recognition community respectively for evaluating Speech Recognizers and Language Models. Early in the series of speech recognition campaigns (Pallett, 2003), NIST provided the

SCLITE (Jurafsky and Martin, 2000) standard evaluation package for computing Word Error Rate. In parallel the creation of specific agencies like the Linguistic Data Consortium in 1992 (Liberman and Cieri, 1998b) and ELRA/ELDA in 1995 (Choukri and Nilsson, 1998), to work as both as repositories of large sized annotated corpora and resource creators, contributed to the development of annotation standards in the natural language processing community.

In that sense their impact on the field is comparable to the release of famous public resources like the Brown corpus (Francis *et al.*, 1979), WordNet (Miller, 1990), the PennTreebank (Marcus *et al.*, 1993) or Propbank (Kingsbury and Palmer, 2002). An example of such contribution are the annotation graphs (Cieri and Bird, 2001) of the Linguistic Data Consortium, which provide a formal framework for representing linguistic annotations of time series data. The formalism offers a logical layer for annotation systems, which is independent from file formats, coding schemes or user interfaces. The impact that language processing evaluation activities had on standards was not limited to providing normalized resources and annotation formats to the NLP community. There was a time when the work of the EAGLES (King *et al.*, 1996) working group on evaluation also influenced the defintion of quality criteria for software proposed by ISO 9126; for an account of the interplay between EAGLE and ISO see (Hovy *et al.*, 2002) and, also in this issue, the article of A. Popescu-Belis (pages 67-91).

But the relationship between standard and evaluation is a two-way relationship, since evaluation can also benefit from the existence of standards for representing and processing language data. On that score, the work of the different subcommittees of ISO-TC37-SC4[8] (Ide and Romary, 2007) whose objective is to prepare various standards by specifying principles and methods for creating, coding, processing and managing language resources, such as written corpora, lexical corpora, speech corpora, dictionary compiling and classification schemes, will be of great help to future evaluation campaign by facilitating interoperability between the various approaches.

8. About this special issue

Of course a special issue on evaluation had to start with an article about evaluation of machine translation (MT), even more so with the increasing interest that MT has attracted recently. In this issue, it is Hervé Blanchon and Christian Boitet who give an account of the history of MT evaluation on page 33, with the intention of showing that MT evaluation should be task based instead of corpus based and should consider only operational systems. The authors distinguish external evaluation methods, based on language output quality and usage performance gain, from internal methods based on system architecture and langugage, document or task portability. For them the current success of statistical oriented methods for MT evaluation comes from the fact that these models have been successfully used in speech recognition.

[8] http://www.tc37sc4.org

Most of the previous MT evaluations were corpus based with little objective judgement. What the authors regret most is that many of the previous evaluations have been done without any real application context and have been carried out on isolated sentences and not whole documents. They would also prefer to see more subjective quality assessment that objective ones, since they argue that when subjective quality is used, the results of manual annotation do not correlate well, for instance, with the results obtained with the BLEU measure. They argue in favor of replacing corpus based evaluation by an external evaluation with evaluation measures linked to the task performed by operational systems adapted to their context of use, in order to measure the usability of the system perceived by the users.

The history of MT evaluation is then presented, beginning with ALPAC, the Japanese JEIDA, the European projects EAGLES, ISLE, the FEMTI proposal, and the 3 campaigns of NESPOLE. Then a critial analysis of the current evaluation methods is done, starting with subjective evaluation criteria (fluidity and adequacy), before addressing BLEU and the other reference based measures. In this section, BLEU is shown to be more a measure of text similarity than a measure of translation quality. Then the authors investigate the recent NIST GALE HTER measure which assesses the quality of a translation through the time spent by a human to correct it. From their analysis the authors offer some proposals for MT evaluation, separately for text and speech data, arguing that evaluation should essentially be done in the context of use of an operational system and should take into account the gain obtained in the average effective translation time.

In the second article of this special issue, we find a general discussion by Andreï Popescu-Belis about external evaluation metrics and corpus based evaluation methods. After investiguating the role of evaluation in NLP, particularly where science and technology meet, the author addresses both the issue of the type of evaluation, and of the size, quality and language coverage of the reference data for corpus based evaluation. To structure his argumentation, A. Popescu uses a broad classification scheme of the various NLP systems based on the types of their input and output. He then considers the methodology one uses in general to produce reference data, with a focus on the question of the kappa, the famouswell-known inter-annotator agreement measure. He explains why kappa is needed to smooth the variations of interpretation proper to human language and advocates its use over the f-measure as an evaluation metric. His articles ends with a discussion on the role of evaluation in the life-cycle of NLP applications.

Sylwia Ozdowska looks at word alignment evaluation in the third article through 3 experiments with the ALIBI alignment system: the first experiment is based on system outputs assessment by human evaluators, the second uses a reference multicorpus and the third one standard public reference data. Before presenting ALIBI (a subsentential alignment system combining analogy on parsing dependencies completed with textual statistics) the author reviews the French ARCADE-I and the American HLT/NAACL-2003, two previous word alignment evaluation campaigns. Three French-English aligned corpora are used in the experiments reported, INRA (agronomy), JOC (questions to the European Commission) and HANSARD (Canadian parliament debates). For the first evaluation experiment, the cases of alignment rule application are checked by hand and precision is computed for each rule. For quantitative evaluation, the translation word pairs and the syntactic information used by the aligner are categorized to define error classes. For the multicorpus evaluation, the author built a reference from several sources and validated the alignment by ensuring that all the annotated data displayed a sufficient value for the kappa statistics, a measure correlated to the inter-annotator agreement. For the multicorpus experiment, the performance of ALIBI was compared with a baseline produced with GIZA++. The HLTNAACL-2003 data set has also been used to provide a performance comparison with other alignment systems, after a necessary adaptation of the word segmentation, which leads the author to discuss the impact on the performance measure of considering in the alignment multi-word expressions or not.

SIMDIAL is the object of the fourth article of this issue. It is the latest achievement in France for SLDS evaluation. Joseph Allemandou, Laurent Charnay, Laurence Devillers, Muriel Lauvergne, and Joseph Mariani first recall why SLDS are among the most difficult NLP systems to evaluate, addressing in particular the issue of genericity, before giving an account of the past SLDS evaluation campaigns. Then they explain their choices of methodology about the following issues: black versus white box evaluation, real versus simulated user, and deterministic versus stochastic simulation. A detailed presentation of SIMDIAL follows. There we learn that SIMDIAL is a black box quantitative evaluation paradigm based on dialog interaction; it can simulate a user and can propose evaluation diagnostics which do not require having access to the SLDS log files. The user dialog turns are generated from an annotated corpus and may include perturbation phenomena classic in spontaneous speech. With SIMDIAL the evaluation is based both on the task completion rate and on the number of dialog turns exchanged with the system. The interaction manager and the dialog model with its six kinds of dialog acts are presented along with the dialog fturn representation formalism. Then the authors provide details of the semantic interpretation mechanism which can be carried out in two modes: automatic or semi-automatic and they also explain how an evaluation diagnostic is automatically produced. They conclude by providing a very detailed account of the application of the paradigm to a resturant information service application.

With the fifth article of this issue, by Jette Viethen and Robert Dale, we look at language generation, a domain that stands apart because, contrary to many other areas of NLP, it has seen the deployment of comparative evaluation campaigns. The authors build their argumentation from an evaluation they did for a task of generating referring expressions in a controlled environment made of a grid of 4x4 locations differentiated by colour and position. They evaluated three different generation algorithms against a corpus of human generated refering expressions, surprisingly a procedure rarely used in the past, as the authors remark. From this experiment, they discuss 4 issues essential for natural language generation evaluation :

1) the influence and the importance that the choice of the knowledge base and its representation (input language) has on the language generation process,

2) the difficulty of establishing a gold standard because of the inherent variability of the output,

3) the numeric performance assessment of natural language generation systems and the issue of defining a preference sorting for the input features and the ouput realizations,

4) the granularity and the domain specificity of the evaluation subtask.

Automatic summaries are addressed in the sixth article of this issue. In order to propose an alternative to the use of the Anglo-Saxon terminology and at the same time to clarify the meaning of terms frequently found in the literature of summary evaluation, Marie-Josée Goulet proposes first a French terminology for the domain. It is organized around three themes : document, summary and actors. Then she makes an extensive review of the literature to collect the various parameters used to describe the evaluation experiments, from which she extracts the main trends of the field. The parameters described concern :

- the source text,
- the automatic summaries being evaluated,
- and the reference summaries.

Of course, the last point about reference data could hardly be discussed without considering the influence that the people who wrote the summaries have on the result of the evaluation and the issue of inter-annotator agreement.

The seventh and last article of this special issue is about evaluation of prosodic phrase break prediction. In their paper, Claire Brierley and Eric Atwell look at the gold standard used for evaluating systems that try to predict the location of prosodic syntactic boundaries, i.e. locations in text where a native speaker would provide a prosodic cue to indicate the end of a syntactic chunk in a natural way. After presenting the task whose main application field is text to speech and the evaluation methodology generally used, the authors review the different machine learning methods applied to phrase breaks prediction, distinguishing rule based methods from statistical ones. Then they present the evaluation measures which for prosodic phrase break prediction are divided in two groups, on the one hand the measures derived from accuracy/error, on the other hand the measures based on precision/recall. The problem of defining a gold standard and achieving a sufficient level of inter-annotator agreement to establish a common reference is addressed next. After this the authors look at the relationship that prosodic phrase break prediction has with respect to parsing and Part Of Speech tagging.

9. Conclusion

Having a special issue about evaluation in NLP is a demonstration of a trend reversal, because 20 years ago the issue of evaluation was controversial in the field. At that time, a majority of actors were not convinced that the benefits outweighed the cost. Since that time, evaluation has gained its legitimacy as a mean to make progress in science and technology, to foster synergy and to support the development of NLP by showing that it can provide applicable solutions for the challenges offered by the processing of the ever growing amount of language data that the development of the Information Society implies. Now that the attrition of evaluation activities in NLP is not to be feared anymore, the next challenges that the evaluation community will face will be the harmonization of the wide range of practices that exist across the various subfields of NLP and across the world, the bridging of the gap between user oriented evaluation and technology evaluation and finally the discovery of new ways to apply evaluation to language generation oriented tasks which up to now have received much less attention than analysis oriented ones.

10. References

Abeillé A., "Analyseurs syntaxiques du français", *Bulletin Semestriel de l'Association pour le Traitement Automatique des Langues*, vol. 32, n˚ 21, p. 107-120, 1991.

Abeillé A., Clément L., Kinyon A., "Building a Treebank for French", *Proceedings of the 2nd International Conference on Language Ressources and Evaluation (LREC)*, Athènes, Grèce, p. 1251-1254, 2000.

Adda G., Lecomte J., Mariani J., Paroubek P., Rajman" M., "The GRACE French Part-of-Speech Tagging Evaluation Task", *in Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, vol. 1, ELDA, Granada, p. 433-441, May, 1998.

Agirre E., Màrquez L., Wicentowski R. (eds), *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Association for Computational Linguistics, Prague, Czech Republic, June, 2007.

Agosti M., Nunzio G. M. D., Ferro N., Harman D., Peters C., *Proceedings of the 11th Conference on Research and Advanced Technology for Digital Libraries*, vol. 4675 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin - Heidelberg, chapter The Future of Large-Scale Evaluation Campaigns for Information Retrieval in Europe, p. 509-512, 2007. ISBN 3540748504, 9783540748502.

Antoine J.-Y., Bousquet-Vernhettes C., Jerome Goulian M. Zakaria Kurdi., Rosset S., Vigouroux N., Villaneau J., "Predictive and objective evaluation of speech understanding: the "challenge" evaluation campaign of the I3 speech workgroup of the French CNRS", *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, ELDA, Las Palmas de Gran Canaria, p. 529-535, May, 2002.

Aston G., Burnard L., *The BNC handbook: exploring the British National Corpus with SARA*, Edinburgh Textbooks in Empirical Linguistics, Edinburgh University Press, Edinburgh, 1998. ISBN 0 7486 1054 5 / ISBN 0 7486 1055 3.

Banko M., Brill E., “Scaling to Very Very Large Corpora for Natural Language Disambiguation”, *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Toulouse, France, p. 26-33, July, 2001.

Blache P., Morin J., “Une grille d’évaluation pour les analyseurs syntaxiques”, *Acte de l’atelier sur l’Evaluation des Analyseurs Syntaxiques dans les actes de la 10^{e} conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*, Batz-sur-Mer, juin, 2003.

Bonneau-Maynard H., Ayache C., Bechet F., Denis A., Kuhn A., Lefevre F., Mostefa D., Quignard M., Rosset S., Servan C., Villaneau J., “Results of the French Evalda-Media evaluation campaign for literal understanding”, *In proceedings of the 5th International Conference on Language Resources and Evaluation*, ELDA, Genoa, p. 2054-2059, May, 2006.

Brant S., Dipper S., Hansen S., Lezius W., Simth G., “The TIGER treebank”, *Proceedings of the 1st Workshop on Treebank and Linguistics Thories (TLT)*, Sozopol, Bulgarie, 2002.

Braschler M., Peters C., *Advances in Cross-Language Information Retrieval*, vol. 2785/2003 of *Lecture Notes in Computer Science*, Springer, Berlin / Heidelberg, chapter CLEF 2002 Methodology and Metrics, p. 512-528, February, 2003. ISBN 978-3-540-40830-7.

Burger J., Palmer D., Hirschman L., “Named Entity scoring for speech input”, *Proceedings of the 17th international conference on Computational linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, p. 201-205, 1998.

Chen S., Beeferman D., Rosenfeld R., “Evaluation metrics for language models”, *In DARPA Broadcast News Transcription and Understanding Workshop*, p. 275-280, 1998.

Chibout K., Mariani J., Masson N., Néel F. (eds), *Ressources et évaluation en ingénierie des langues*, Universités francophones, DeBoeck, 2000. ISBN 2-8011-1258-5.

Choukri K., Nilsson M., “The European Language Resources Association”, *In proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, ELDA, Granada, p. 153-158, May, 1998.

Cieri C., Bird S., “Annotation Graphs and Servers and Multi-Modal Resources: Infrastructure for Interdisciplinary Education, Research and Development”, *Proceedings of the ACL Workshop on Sharing Tools and Resources for Research and Education*, Toulouse, France, p. 23-30, July, 2001.

Cleverdon C., “The ASLIB Cranfield research project on the comparative efficiency of indexing systems”, *ASLIB Proceedings*, vol. 12, p. 421-431, 1960. ISSN: 0001-253X / DOI: 10.1108/eb049778.

Cole R., Mariani J., Uszkoreit H., Zaenen A., Zue V., Survey of the state of the art in human language technology, Technical report, Center for Spoken Language Understanding CSLU, Carnegie Mellon University, Pittsburgh, 1996. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html.

Dybkjæer L., Bernsen N. O., Carlson R., Chase L., Dahlbäck N., Failenschmid K., Heid U., Heisterkamp P., Jönsson A., Kamp H., Karlsson I., v. Kuppevelt J., Lamel L., Paroubek P., Williams D., “The DISC approach to spoken language systems development and evaluation”, *Actes de la 1ère International Conference on Language Resources and Evaluation (LREC98)*, vol. 1, ELRA, Granada, Spain, p. 185-189, May, 1998.

Dybkjæer L., Ole Bernsen N., Minker W., “Evaluation and usability of multimodal spoken language dialogue systems”, *Speech Communication*, vol. 43, n° 1-2, p. 33-54, June, 2004. doi:10.1016/j.specom.2004.02.001.

Edmonds P., Kilgarriff A., “Special issue based on Senseval-2”, *Journal of Natural Language Engineering*, January, 2003.

Ellis D., *New Horizons in Information Retrieval*, The Library Association, London, 1990.

Ellis D., “The Physical and Cognitive Paradigm in Information Retrieval Research”, *Journal of Documentation*, vol. 48, n° 1, p. 45-64, 1992. ISSN:0022-0418, DOI:10.1108/eb026889.

Fisher K., Erdelez S., McKechnie L. (eds), *Theories of Information Behaviour*, ASIST Monograph Series, Information Today, Medford, NJ, 2005.

Francis W. N., , Kuc̀era H., *Manual of Information to Accompany a Standard Sample of Present-day Edited American English, for Use with Digital Computers*, original ed. 1964, revised 1971, revised and augmented 1979 edn, Providence, R.I.: Department of Linguistics, Brown University. 1979.

Grishman R., Sundheim B., “Message Understanding Conference- 6: A Brief History”, *COLING*, p. 466-471, 1996.

Harman D., “The DARPA TIPSTER project”, *ACM SIGIR Forum*, vol. 26, n° 2, p. 26-28, 1992. ISSN:0163-5840.

Harter S. P., “Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness”, *Journal of the American Society for Information Science and Technology (JASIST)*, vol. 47, p. 37-49, January, 1996.

Hauser R., “Results of the 1. Morpholympics”, *LDV-FORUM*, June, 1994. ISSN 0172-9926.

Hirschman L., “Language understanding evaluations: lessons learned from MUC and ATIS”, *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, Grenade, Espagne, p. 117-122, 1998.

Hirschman L., Bates M., Dahl D., Fisher W., Garofolo J., Hunicke-Smith K., Pallett D., Pao C., Price P., Rudnicky A., “Multi-Site Data Collection for a Spoken Language Corpus”, *Proceedings of ICSLP-92*, Banff, Canada, p. 9-14, October, 1992. ISBN:1-55860-272-0.

Hovy E., King M., Popescu-Belis A., “An Introduction to MT Evaluation”, *in* M. King (ed.), *Workbook of the LREC 2002 Workshop on Machine Translation Evaluation: Human Evaluators Meet Automated Metrics*, Las Palmas, p. 1-6, May, 2002.

Ide N., Romary L., *Evaluation of Text and Speech Systems*, vol. 36 of *Text, Speech and Language Technology*, Kluwer Academic Publisher, chapter Towards International Standards for Language Resources, p. 263-284, 2007. ISBN-10: 1-4020-5815-2, ISBN-13: 978-1-4020-5815-8.

Jouet J., “Retour critique sur la sociologie des usages”, *Réseaux*, 2000.

Jurafsky D., Martin J., *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall, 2000. ISBN 0130950696.

Kaplan F., *La naissance d’une langue chez les robots*, Hermès Science, 2001. ISBN 2-7462-0262.

King M., “When is the next Alpac report due?”, *in Proceedings of the 22nd annual meeting on Association for Computational Linguistics table of contents*, Association for Computational Linguistics Morristown, NJ, USA, Stanford, California, p. 352-353, 1984. DOI: 10.3115/980491.980563.

King M., Maegaard B., Schütz J., des Tombes L., Bech A., Neville A., Arppe A., Balkan L., Brace C., Bunt H., Carlson L., Douglas S., Höge M., Krauwer S., Manzi S., Mazzi

C., Sieleman A. J., Steenbakkers R., *EAGLES Evaluation of Natural Language Processing Systems*, Center for Sprogteknologi, Cophenhaguen, october, 1996. ISBN 87-90708-00-8.

Kingsbury P., Palmer M., “From Treebank to Propbank”, ELDA, Las Palmas, p. 1989-1993, May, 2002.

Lamel L., Minker W., Paroubek P., “Towards Best Practice in the Development and Evaluation of Speech Recognition Components of a Spoken Language Dialogue System”, *Revue Natural Language Engineering*, vol. 6, n˚ 3, p. 305-322, october, 2000.

Lavid J., “To the memory of John Sinclair, Professor of Modern English Language”, *Estudios Ingleses de la Universidad Complutense*, vol. 15, p. 9-12, 2007. ISSN: 1133-0392.

Lehmann S., Estival D., Oepen S., “TSNLP Des jeux de phrases-test pour l’evaluation d’applications dans le domaine du TALN”, *Actes de la conférence sur le Traitement Automatique de la Langue Naturelle (TALN 1996)*, Marseille, May, 1996.

Liberman M., Cieri C., “The Creation, Distribution and Use of Linguistic Data”, *In proceedings of the First International Conference on Language Resources and Evaluation*, ELRA, Granada, Spain, May, 1998a.

Liberman M., Cieri C., “The Creation, Distribution and Use of Linguistic Data: the case of the Linguistic Data Consortium”, *In proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, ELDA, Granada, p. 159-164, May, 1998b.

Magnini B., Cappelli A. (eds), *Evalita 2007: Evaluating Natural Language Tools for Italian*, vol. IV n°2, Associazione Italiana Intelligenza Artificiale (AI*IA), Roma, June, 2007. ISSN 1724-8035.

Manning C. D., Schütze H., *Foundation of Statistical Natural Language Processing*, 5ème edn, Massachusetts institute of Technology Press, 2002.

Mapelli V., Nava M., Surcin S., Mostefa D., Choukri K., “Technolangue: A Permanent Evaluation and Information Infrastructure”, *In proceedings of the 4th international Conference on Language Resources and Evaluation (LREC)*, vol. 2, ELDA, Lisboa, Portugal, p. 381-384, May, 2004.

Marcus M. P., Marcinkiewicz M. A., Santorini B., “Building a Large Annotated Corpus of English: The Penn Treebank”, *Computational Linguistics*, 1993.

Mariani J., “Developing Language Technologies with the Support of Language Resources and Evaluation Programs”, *Language Resources and Evaluation*, vol. 39, n˚ 1, p. 35-44, september, 2005. DOI 10.1007/s10579-005-2694-3.

Mariani J., Paroubek P., “Human Language Technologies Evaluation in the European Framework”, *Proc. of the DARPA Broadcast News Workshop*, Morgan Kaufmann, Herndon, VA, p. 237-242, February, 1999.

Olfati-Saber R., Fax A., Murray R., “Consensus and Cooperation in Networked Multi-Agent Systems”, *Proceedings of the IEEE*, vol. 95, n˚ 1, p. 215-232, January, 2007. DOI 10.1109/JPROC.2006.887293.

Mey M. D., “The Cognitive Viewpoint: Its development and its Scope”, *Proceedings of the International Workshop on the Cognitive Viewpoint*, Ghent, p. 16-32, 1977.

Miller A., “Wordnet: An on-line lexical database. International journal of Lexicography”, *International journal of Lexicography*, vol. 3, n˚ 4, p. 235-312, 1990.

Mitkov R. (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford Handbooks in Linguistics, Oxford University Press, Januray, 2005. ISBN-13: 978-0-19-927634-9.

Oepen S., Netter K., Klein J., “Test Suites for Natural Language Processing”, *CSLI Lecture Notes*, Center for the Study of Language and Information, 1996.

Paek T., “Empirical Methods for Evaluating Dialog Systems”, *SIGdial Workshop on Discourse and Dialogue*, 2001.

Pallett D., “A look at NIST’S benchmark ASR tests: past, present, and future”, *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, IEE, Virgin Islands, USA, p. 483-488, November, 2003. ISBN:0-7803-7980-2 / DOI:10.1109/ASRU.2003.1318488.

Palmer M., Finin T., “Workshop on the Evaluation of Natural Language Processing Systems”, *Computational Linguistics*, vol. 16, n° 3, p. 175-181, 1990.

Papineni K., Roukos S., Ward T., Zhu W.-J., “Bleu: a Method for Automatic Evaluation of Machine Translation”, *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, p. 311-318, July, 2002.

Paroubek P., *Evaluation of Text and Speech Systems*, vol. 36 of *Text, Speech and Language Technology*, Kluwer Academic Publisher, chapter Evaluating Part Of Speech Tagging and Parsing, p. 97-116, 2007. ISBN-10: 1-4020-5815-2, ISBN-13: 978-1-4020-5815-8.

Paul M., “Overview of the IWSLT 2006 Evaluation Campaign”, *Proceedings of the International Workshop on Spoken Language Translation - Evaluatio Campaign on Spoken Language Translation*, ATR, Kyoto, p. 1-15, november, 2006. http://www.slt.atr.jp/IWSLT2006/archives/2005/11/proceedings.html.

Pierrel J.-M., “Un ensemble de ressources de référence pour l’étude du français : TLFI, FRANTEXT et le logiciel STELLA”, *Revue québécoise de linguistique*, vol. 32, n° 1, p. 155-176, 2003.

S. Nirenburg H. S., Wilks Y. (eds), *Readings in Macine Translation*, MIT Press, Cambridge, Massachusset, p. 131-135, 2003. ISBN-10: 0-262-14074-8, ISBN-13: 978-0-262-14074-4, http://www.hutchinsweb.me.uk/ALPAC-1996.pdf.

Sampson G., *English for the Computer: The SUSANNE Corpus and analytic scheme*, Clarendon Press, Oxford, 1995. ISBN 0-19-824023-6.

Santos D., Cardoso N., *A Golden Resource for Named Entity Recognition in Portuguese*, vol. 3960 of *Lecture Notes in Computer Science*, Springer, Berlin / Heidelberg, p. 69-79, 2006. ISBN 978-3-540-34045-4, DOI 10.1007/11751984_8.

Santos D., Costa L., Rocha P., “Cooperatively evaluating Portuguese morphology”, *in* N. J. Mamede, J. Baptista, I. Trancoso, M. das Graças Volpe Nunes (eds), *in Proceedings of the 6th international workshop on Computational Processing of the Portuguese Language (PROPOR)*, Springer-Verlag, Faro, p. 259-266, June, 2003. http://www.linguateca.pt/Diana/download/SantosCostaRochaPROPOR2003.pdf.

Saracevic T., “Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II: Nature and Manifestations of Relevance”, *Journal of the American Society for Information Science and Technology (JASIST)*, vol. 58, p. 1915-1933, November, 2007.

Schamber L., “Relevance and Information Behavior”, *Annual Review of Information Science and Technology*, vol. 29, p. 3-48, 1994.

Sparck Jones, *User Models in Dialog Systems*, Springer-Verlag, Berlin, chapter Realism about User Modeling, p. 341-363, 1990.

Sparck Jones, Galliers J. R., *Evaluating Natural Language Processing Systems. An Analysis and Review*, n˚ 1083 in *Lecture Notes in Artificial Intelligence*, Springer-Verlag, Berlin, 1996.

Sparck Jones K., "Automatic language and information processing: rethinking evaluation", *Natural Language Engineering*, vol. 7, n˚ 1, p. 29-46, 2001. Cambridge University Press.

Sperber D., Wilson D., *La Pertinence*, Communication et cognition, Les Editions de Minuit, Paris, 1989.

Voorhees E. M., "The philosophy of information retrieval evaluation", *In Proceedings of the The Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, Springer-Verlag, p. 355-370, 2002.

Voorhees E. M., Harman D. K. (eds), *TREC: Experiment and Evaluation in Information Retrieval*, Digital libraries and electronic publishing series, william y. arms edn, The MIT Press, Cambridge, MA, 2005. ISBN 0-262-22073-3.

Walker M. A., Passonneau R., Boland J. E., "Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems", *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Toulouse, France, p. 515-522, July, 2001a.

Walker M., Aberdeen J., Boland J., Bratt E., Garofolo J., Hirschman L., Le A., Lee S., Narayanan S., Papineni K., Pellom B., Polifroni J., Potamianos A., Prabhu P., Rudnicky A., Sanders G., Seneff S., Stallard D., Whittaker S., "DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection", *In proceedings of the 7th European Conference on Speech Processing (EUROSPEECH)*, Aalborg, Denmark, p. 1371-1374, September, 2001b. http:www.isca-speech.orgarchiveeurospeech_2001e01_1371.html.

Walker M., Kamm C., Litman D., "Towards developing general models of usability with PARADISE", *Natural Language Engineering archive*, vol. 6, n˚ 3-4, p. 363-377, September, 2000. ISSN:1351-3249 / DOI 10.1017/S1351324900002503.

Wayne C. L., "A Snapshot of two Darpa Speech and Natural Language Programs", *Proceedings of the Speech and Natural Language Workshop*, Pacific Grove, California, p. 403-404, February, 1991.