The PEACE SLDS understanding evaluation paradigm of the French MEDIA campaign

Laurence Devillers, Hélène Maynard, Patrick Paroubek, Sophie Rosset
LIMSI-CNRS
Bt 508 University of Paris XI - BP 133 F-91403 ORSAY Cedex, France
{devil,hbm,pap,rosset}@limsi.fr

Abstract

This paper presents a paradigm for evaluating the context-sensitive understanding capability of any spoken language dialog system: PEACE (French acronym for *Paradigme d'Evaluation Automatique de la Compréhension hors et En-contexte*). This paradigm will be the basis of the French Technolangue MEDIA project, in which dialog systems from various academic and industrial sites will be tested in an evaluation campaign coordinated by ELRA/ELDA (over the next two years). Despite previous efforts such as EAGLES, DISC, AUPELF ARCB2 or the ongoing American DARPA COMMUNICATOR project, the spoken dialog community still lacks common reference tasks and widely agreed upon methods for comparing and diagnosing systems and techniques. Automatic solutions are nowadays being sought both to make possible the comparison of different approaches by means of reliable indicators with generic evaluation methodologies and also to reduce system development costs. However achieving independence from both the dialog system and the task performed seems to be more and more a utopia. Most of the evaluations have up to now either tackled the system as a whole, or based the measurements on dialog-context-free information. The PEACE proposal aims at bypassing some of these shortcomings by extracting, from real dialog corpora, test sets that synthesize contextual information.

1 Introduction

Generally speaking common reference tasks (Whittaker et al., 2002) and methods to compare and diagnose spoken language dialog systems (SLDS) and spoken dialog techniques are lacking despite previous efforts futher discussed in the next section such as EAGLES, DISC, AUPELF ARCB2 or the ongoing American project DARPA COMMUNICATOR. Without an objective assessment of dialog systems, it is difficult to reuse previous work and to advance theories. The assessment of a dialog system is complex in part to the high integration factor and tight coupling between the various modules present in any SLDS, for which unfortunately today, no common accepted reference architecture exists. Nevertheless, a major problem remains the dynamic nature of dialog. Consequently to these shortcomings, researchers are often unable to provide principled design and system capabilities for technology transfer. In other research areas, such as speech recognition and information retrieval, common reference tasks have been highly effective in sharing research costs and efforts. A similar development is highly needed in the dialog community.

In this contribution which addresses only a part of the SLDS evaluation problem, a paradigm for

evaluating the context-sensitive understanding capability of any spoken language dialog system is proposed. PEACE (Devillers et al., 2002a) described in section 3, is based on test sets extracted from real corpora, and has three main aspects: it is generic, contextual and it offers diagnostic capabilities. Here genericity is envisaged in a context of information dialogs access. The diagnostic aspect is important in order to determine the different qualities of the systems under test. The contextual aspect of evaluation is a crucial point since dialog is dynamic by nature. We propose to simulate/synthesize the contextual information. The PEACE paradigm will be tested in the French Technolangue MEDIA project and will serve as basis in the comparison and diagnostic evaluation of systems presented by various academic and industrial sites (section 4). ELRA/ELDA is the coordinator of the larger scope evaluation campaign EVALDA, which includes the MEDIA campaign that began in January 2003.

2 Overview of SLDS evaluation

Without an attempt to be exhaustive, we overview some recent efforts for evaluation of SLDS.

The objective of the European DISC project was to write the best-practice guidelines for SLDS development and evaluation of its time. DISC has collected a systematic list of bottom-up evaluation criteria, each corresponding to a partially ordered list of properties likely to be encountered in any SLDS. This properties are positioned on a grid defining an SLDS abstract architecture and relate to various phases of the generic DISC SLDS development life-cycle (Dybkjær and al., 1998). They are complemented by a standard evaluation pattern made of 10 generic questions (e.g. “Which symptoms need to be observed?”) which has been instantiated for all the evaluation criteria. If the DISC results are quite extensive and presented in an homogeneous way, they do not provide a direct answer to the question of SLDS evaluation. Its contribution lies more at the specification level. Although the approach and the goals of the European EAGLES project were different, one could forward the same remark about the results of the speech evaluation work group (D. Gibbon, 1997). In (Fraser, 1998), one find a set of evaluation criteria for voice oriented products and services, organized in four broad categories.: 1) voice command, 2) document generation, 3) phone services 4) other.

To the best of our knowledge, the MADCOW (Multi Site Data COllection Working group) coordination group set up in the USA by ARPA in the context of the ATIS (*Air Travel Information Services*) task to collect corpora, was the first to propose a common infrastructure for SLDS automatic evaluation (MADCOW, 1992), which also addressed the problem of language understanding evaluation, based on system answer comparison. Unfortunately no direct diagnostic information can be produced, since understanding is appreciated by gauging the distance from the answer to a pair of minimal and a maximal reference answers. In ATIS, the protocol was only been applied to context free sentences. Up to now it has been one of the most used by the community since it is relatively objective and generic because it relies on counts of explicit information and allows for a certain variation in the answers. On the other hand, the method displays a bias toward silence and does not give the means to appreciate error severity.

In ARISE (*Automatic Railway Information Systems for Europe*) (Lamel, 1998), a corpus of roughly 10,000 calls has been used in conjunction with user debriefing questionnaire analysis to diagnose different versions of a phone information server. The hand-tagging objective measures of the corpus include understanding error counts (glass box methodology). Although it provides fine grained diagnostic information, this procedure cannot be easily generalized since it requires hand-annotated corpus and access to the internal representation of the system.

Two metrics have been developped at MIT (Glass et al., 2000): the *Query Density* (QD) and the *Concept Efficiency* (CE), which measure respectively over the course of a dialogue: the mean number of new concepts introduced per user query, and the number of turns necessary for each concept to be understood by the system. Concepts are generated automatically for each utterance with a parsable orthographic transcription as a series of keyword-value pairs. The higher the

QD, the more effectively a user is able to communicate information to the system. The CE is an indicator of recognition or understanding errors; the higher it is, the fewer times a user needs to repeat himself. These metrics were evaluated on single systems (JUPITER and and MERCURY); to compare different systems of the same type, one would need a common ontology. In (Glass et al., 2000), the authors believe that CE should be related to user frustation, but to show it they would need to use the PARADISE framework.

PARADISE (Walker et al., 1998) can be seen as a sort of meta-paradigm which correlates objective and subjective measurements. Its grounding hypothesis states that the goal of any SLDS is to achieve user-satisfaction, which in turn can be predicted through task success and various interaction costs. With the help of the kappa coefficient (Carletta, 1996) proposes to represent the dialog success independently from the task intrinsic complexity, thus opening the way to task generic comparative evaluation. PARADISE has been tested in the COMMUNICATOR project (Walker et al., 2001) with 9 systems working on the same task over different databases. With four basic measures (e.g. task completion) the protocol has been able to predict 37% of user satisfaction variation, and 42% with the help of a few extra measurements on dialog acts and subtasks. One critic, one can make about PARADISE concern its cost (real user tests are costly) and the use of subjective assessment.

The adaption of the DQR text understanding evaluation methodology (Sabatier et al., 2000) to speech resulted in a generic and qualitative procedure. Each element of its test set holds three parts, the *Declaration* to define the context, a *Question* which bears on point present in the context and the *Response*. The test set is organized through seven levels of test, from basic explicit understanding to semantic interpretation and reply pertinence assessment. This protocol is task and system generic but test set construction is not straightforward and the bias introduced by the wording of the question is difficult to assess.

Recently the GDR-13 work group of CNRS on spoken dialog understanding, has proposed an evaluation methodology for literal understanding. According to (Antoine and al., 2002), DEFI tries to remedy two important weaknesses of the MADCOW methodology, namely the lack of genericity and the lack of diagnostic information, by crafting system specific test sets from a primary set of enunciations representative of the task (provided by the developers). Secondary enunciations are then derived from the primary ones in order to exhibit particular language phenomena. Afterwards, the systems are evaluated by their developers using specific test set and their own metrics. The various results can be mapped over a generic abstract architecture for comparison (although this mapping is still unspecified at the time of writing). DEFI has already been used in one evaluation campaign, with 5 systems presented by 4 laboratories. (Antoine and al., 2002) has reported the following weaknesses of the protocol: how to control the bias introduced by the derivation of enunciations, how to guaranty that derived enunciation will remain in the task scope (this prevented some system from being evaluated over the complete test set) and finally how to restrict and organize the language phenomena used in the test set.

3 The PEACE paradigm

We first describe the paradigm and relate preliminary experiments with PEACE. This paradigm which is as basement for the MEDIA project will be refined by all the partners and use for an evaluation campaign between seven systems of industrial and academic sites.

3.1 Description

The PEACE paradigm relies on the idea that for database querying tasks, it is possible to define a common semantic representation, onto which all the systems are able to convert their own representation (Moore, 1994). The paradigm based on data extracted from real corpus, includes both literal and contextual understanding test sets. More precisely, it provides:

- the definition of a semantic representation (see 3.1.1),
- the definition of a model for dialogic contexts (see 3.1.2),

- the definition and typology of linguistic phenomena and dialogic functions used to selectively diagnoze the system language capabilities (anaphora resolution, constraints relaxation, etc.) (see 3.1.3),
- a data structuring method. The format of the annotated data will be adapted to language resource standard annotations implemented (see 3.1.4),
- and evaluation metrics with the corresponding evaluation tool (see 3.1.5).

3.1.1 Generic semantic representation

The difficulty of choosing a semantic representation lies in finding a complete and simple representation of a user utterance meaning in a unified format. A frame Attribute Value Representation (AVR) has been chosen, allowing a fast and reliable annotation. The values are either numeric units, proper names, or semantic classes, that group together lexical units which are synonyms for the task. The order of the (attribute, value) pairs in the semantic representation matches their respective position in the utterance. A modal information (positive (+) and negative(-)) is also assigned to each (attribute, value) pair. The semantic representation of an utterance consists then in a list of triplets of the form (mode, attribute, normalized value). An example is given in figure 1. In order to take into account for long-time dependencies or to allow multiple referenced objects, the semantic representation may be enriched by adding a *reference* value to each triplet for the representation of links between 2 attributes of the utterance.

Attributes can grouped into different classes:

- the **database attributes** (the most frequent) correspond to the attributes of the database tables (e.g. `category` for an hotel);
- the **modifier attributes** are associated to the database concepts. Their values are used to modify the database concept interpretation values (e.g. the attribute `category-modifier` with possible values: $>$; $<$, $=$, `Max`, `Min`);
- the **discursive attributes** are introduced to handle various aspects of dialogic interaction (e.g. `command` with values *cancelation*, *correction*, *error specification*..., or *response* with values *yes* or *no*);
- the **argument attribute** which represents the topic at the focus of the utterance.

User Query	*c'est pas Paris c'est Passy* **it is not Paris it is Passy**
(LU) AVR	(-, place, Paris) (+, place, Passy)

Figure 1: Example of a semantic representation of an utterance with positive and negative information for the ARISE task. `Place` is an database attribute,`Paris` and `Passy` are values and +/- modal markers.

When dealing with information retrieval applications, defining the database and modifier attributes and the appropriate values can be done in a rather straightforward way. Most of those attributes are derived directly from the information stored in the database. Furthermore, most of the discursive attributes are domain-independent. Some database attributes remain unchanged across many tasks, such as those dealing with dates or prices.

This semantic representation has been used at LIMSI for PARIS-SITI TASK (touristic information) and ARISE TASK (traintable information) both with triplet representation. More recently in the context of the AMITIES project, quadruplets were used.

3.1.2 Contextual understanding modeling

Contextual understanding evaluation provides information about the capability of the system to take into account the dialog history in order to properly interpret the user query. Contextual understanding evaluation is rarely performed because of the dynamic nature of the dialog make the dialog context depend on the system's dialog strategy.

Nevertheless PEACE proposes a system-independent way to evaluate local contextual interpretation. Given $U_1 ... U_t$ the user interactions, and $S_1 ... S_t$ the answers of the agent or system, the context a time t is a function $f(U_1, S_1, U_2, S_2, ... U_t, S_t)$. In the PEACE paradigm, a *paraphrase of the context* is derived

from the semantic representation (Bonneau-Maynard et al., 2000).

The dialog contexts are extracted from real dialogs in three steps. First, the internal semantic frames representing the dialog contexts are automatically extracted from the log files of the session recordings. Secondly, the semantic frames are converted into AVR format and then hand-corrected to faithfully represent the dialog history. The last step consists in the writing of a sentence for each context (the context paraphrase), which results in the same AVR representation as the one of the dialog context.

Two possibilities may be investigated for building the paraphrase from the internal semantic representation of the dialog context. A rule-based or template-based natural language generation module can be used to automatically produce the paraphrase. The paraphrase can also be obtained by concatenating the sentences preceding the extracted dialog state. In both cases, a manual verification is needed.

3.1.3 A typology of linguistic phenomena and dialogic functions

For dialog system evaluation, it is essential to build test sets randomly extracted from real corpus. For dialog system diagnosis, it is also crucial to build test sets labeled with the linguistic phenomena and dialogic functions. Thus, the capabilities of system's contextual understanding can be assessed for the main linguistic and dialogic difficulties such as, for instance, anaphora or ellipsis resolution.

3.1.4 A data structuring method

Two types of units, one for literal understanding (LU), the other for contextual understanding (CU) are defined. The format of the annotated data will be adapted to language resource standard annotations implemented in XML, e.g. (Geoffrois et al., 2000), (Ide and Romary, 2002).

Each unit is extracted from a real dialog corpus. LU units are composed of the user query, the corresponding audio signal, an automatic transcription obtained with a recognition system, and finally the literal semantic representation of the utterance (see Figure 1). CU units are composed of the dialog context (given by the paraphrase), the user query and the resulting AVR of the user query in the given context (see Figure 2). Those units are also labeled with linguistic and dialogic phenomena.

Context paraphrase	*je voudrais un hôtel 4 étoiles dans le neuvième* **I would like a 4 category hotel in the ninth**

`(LU) AVR`	`(+, argument, hotel)`
	`(+, district, 9)`
	`(+, category, 4)`

User query	*la même catégorie dans un autre arrondissement* **the same category in another district**
`(LU) AVR`	`(+, other, district)` `(+, same, category)`
(CU) AVR	**(+, argument, hotel)** **(-, district, 9)** **(+, category, 4)**

Figure 2: Example of a contextual understanding unit composed of a context paraphrase, a user query and the resulting AVR. AVR of context paraphrase and user query are given in TYPEWRITING MODE. Ellipsis (*"in the ninth"*) and anaphora (*"same category", "another district"*) may be observed.

3.1.5 Evaluation metrics and scoring tool

Common evaluation metrics are essential for analyzing the system capabilities. The scoring tool for AVR comparison is able to compare between two AVR frame representation sets. For evaluation, system outputs translated in AVR format composed one set, the other one contains the AVR references which are manually annotated. Both frame sets have the form of a list of AVRs (fixed length records). Each record is composed of three or four fields (mode, attribute, value, reference). The comparison consists in applying a set of predefined operators each assigned with a cost value. The comparison process looks for operator lists to be applied to the test frame in order to obtain the reference frame that minimizes the final cost value. For a global evaluation, the classical operators from speech evaluation (DELetion, INSertion and SUBstitution) may be used (as used for first two values of `Accuracy` percentage in Table 1).

With our scoring tool the definition of new operators is quite easy. It is then also possible to distinguish between different types of errors by defining specific operators (as used to estimate Topic identification in Table 1), or by using different cost values (for example a substitution is often considered more costly for dialog management).

3.2 Example use of PEACE

In order to validate the evaluation paradigm, a set of approximatively 1,700 literal units and a set of 100 contextual units has been used for the PARIS-SITI task (Bonneau-Maynard and Devillers, 2000). Results for both literal and contextual understanding test sets are given in Table 1. In order to observe the ability of the systems to deal with recognition errors, each literal understanding unit also contains the ASR transcription of the original user utterance. The various measures of understanding accuracy are computed as the ratio between the sum of the number of deleted, inserted and substituted attributes, and the total number of AVR attributes in the test set. The possibility of an automatic evaluation of the LU accuracy and the ability of the scoring tool to point out the errors allowed us to easily improve the literal understanding accuracy from 89.0% to 93.5%. Due to a 26.5% ASR error rate, the LU accuracy goes down from 93.5% to 72% after ASR transcription. The contextual understanding accuracy on the 100 test units is 82.6% on exact transcription. For instance, anaphoric references are relatively well solved, with 80.4% accuracy on the 50 units containing at least one anaphoric reference. For each example, the anaphoric referenced object is generally correctly identified and remaining errors are often due to a bad history constraint management.

3.3 Discussing the PEACE paradigm

The PEACE paradigm enables automatic evaluation of literal and contextual dialog understanding. The evaluation paradigm makes the distinction between different types of errors, allowing a qualitative and diagnostic analysis of the performances of a speech understanding module. Very few evaluation paradigms propose automatic diagnosis of contextual interpretation (Glass et al., 2000). The proposed methodology is based on semi-automatically built reference test sets, and therefore is much more time effective than manual evaluation. Furthermore, it provides reproducible tests.

	#Units	#Attr.	%Acc.	Prec.
LU exact	1 681	3 991	93.5%	0.7
LU ASR .	1 681	3 991	72.0%	1.4
Topic id.	680	833	94.3%	1.6
Modifier id.	323	445	95.7%	1.9
CU exact	100	430	86.8%	3.2
Anaphoric resolution	50	245	84.4%	4.5
Ellipsis resolution	25	106	85.3%	6.7

Table 1: Literal understanding (LU) accuracy on both exact and ASR transcription, and contextual understanding (CU) accuracy. Second column indicates the number of units included in the test set (i.e # of user utterances), third column gives the total number of attributes in the correct AVR test sets. Details, using specific operators, are given for `argument` (topic) and `modifier` identification for LU on exact transcription, and for anaphoric reference and ellipsis resolution for CU. Last column gives the 95% precision of the accuracy estimation (Montacié and Chollet, 1997)

.

Although the semantic representation is task dependent, the example described above shows the feasibility of the paradigm for any dialog system interfacing to a database. Robustness to many linguistic phenomena such as repetitions, hesitations or auto-corrections may be evaluated with this method. XML coding will facilitate the genericity and the reusability of the test sets, by allowing the selection of the dialogic contexts to be studied.

The representation of the dialog context with a single paraphrase, derived from a "flat" structured AVR, may have some limitations in case of long-time dialog dependencies. It does not allow for memorizing all the steps of the dialog. For example, if the speaker says first "*I would like a 2 star hotel*", then "*no I prefer 3 stars*" and finally says "*give me again my first choice*", the CU unit cannot take into account this succession of queries. However, this kind of interaction is rarely observed in dialogue corpora: the user usually repeats the constraint value ("*give me again a 2 star hotel*"). To represent more precisely the

dialog state, the representation of the dialog context should incorporate some meta-information inspired for example from the DAMSL annotation standard [1] (Devillers et al., 2002b).

Another point is the representativity of the test sets. This may be considered as a limitation as far as PEACE paradigm is built on the idea that the test units are extracted from real dialogs. Obviously, the larger the test sets are, the better. A diagnostic evaluation may need a very large test corpora to validate system performance against the wide range of phenomena present in spontaneous dialog.

The ability to automatically diagnose the performances of contextual understanding modules on local difficulties such as ellipsis, negations, anaphoric reference or constraint relaxation is one of the major advantages of the PEACE paradigm, which has not been investigated by other methodologies. This is why it has been chosen for the MEDIA project described in the next section.

4 The MEDIA project

The MEDIA project proposes a paradigm based on a reference task and on test sets extracted from real corpora for evaluating literal and contextual understanding in dialog systems. The PEACE paradigm will serve as basis for the MEDIA project. The consortium is composed of IRIT, LIA, LIMSI, LORIA, VALORIA for the French academic sites and France Telecom R&D and TELIP for the industrial sites. The scientific committee contains representatives of AT&T (USA), Tilburg University (Netherlands), IBM, IMAG, LIUM and VECSYS (France).

The project has four main parts. First, the selection of reference task such as for example a task of web-based travel agency. The reference task has to correspond to a real-life application allowing real user tests. Secondly, multi-level representation such as the semantic representation, the typology of linguistic phenomena and dialogic functions, the dialog context model... will be commonly refined and adapted to the reference task. The third part deals with the recording and labeling of a dialog corpus which will be used for both system adaptation and test set selection. The last part is the organisation of the evaluation campaigns by ELRA/ELDA for the participating sites.

ELRA/ELDA is the coordinator of a larger scope project: EVALDA which includes among others, the MEDIA project. ELDA with VECSYS will provide transcribed and annotated corpora and evaluation tools according to consortium specifications. The recording of 1200 French dialogs (240 speakers, 5 dialogs each, 15k user queries) is planned. Three sets of LU and CU units will be built from this corpus. A large size adaptation set will be used by the participants to adapt their system to the task and the semantic representation. The development set (around 1K LU (resp. CU) units) will be used to validate the evaluation protocole. The size of the test set is planned to be around 3K LU (resp. CU) units. Various approaches are currently used at the participating sites; stochastic or syntactic and semantic rule-based modeling. The project started in January 2003 and will last two years.

5 Conclusion

Assessing the dialog system understanding capabilities requires to evaluate the transition between successive states of the dialog. At least, we must be able to test a sequence of two states at any point in the dialog. The dynamic and interactive nature of the dialog makes construction and reuse of test sets difficult. Furthermore, to evaluate one particular dialog transition, the system has to be put in a particular state corresponding to the original dialog context. The variable describing the dialog state can be composed of complex information such as the current semantic frame (list of triplets (mode,attribute,value) or quadruples (mode, attribute, value, reference)), the dialog history semantic frame and potentially other information like recognition scores, dialog acts, etc.

The PEACE paradigm allows the evaluation of two successive simplified dialog states. It has been successfully tested with test samples focusing on linguistic difficulties of literal and contextual understanding. For these tests, the dialog state is the dialog history semantic frame. The contextual understanding modeling in PEACE is *system independent* since the context is given by a paraphrase

[1] http://www.cs.rochester.edu/research/trains/annotation

of queries. PEACE allows a *diagnostic evaluation* of specific semantic attributes and *particular linguistic phenomena*.

In our opinion, it is crucial for the dialog community to agree on a *common reference task* and reference test sets in order to be able to compare and diagnose dialog systems. Both evaluation with real users and artificial simulation of successive dialog states using test sets extracted from real corpora have to be carried out in parallel. The use of test sets reduces the global cost of dialog system evaluation, moreover such tests are reproducible.

The PEACE protocol will be used as basis for the French Technolangue MEDIA project in a two year evaluation campaign where dialog systems from both academia and industry will be evaluated. In other domains, it could be related with (Hirschman, 2000) propositions for Question Answering evaluation.

References

J.Y. Antoine and al. 2002. Predictive and objective evaluation of speech understanding: the "challenge" evaluation campaign of the i3 speech workgroup of th french cnrs. In *LREC2002*, Spain, May. ELRA.

H. Bonneau-Maynard and L. Devillers. 2000. A framework for evaluating contextual understanding. In *ICSLP*.

H. Bonneau-Maynard, L. Devillers, and S. Rosset. 2000. Predictive performance of dialog systems. In *LREC2000*, volume 1, pages 177–181, Athens, Greece, May. ELRA.

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 2(22):249–254.

R. Winski D. Gibbon, R. Moore. 1997. *Handbook of Standards and Ressources for Spoken Language Ressources*. Mouton de Gruyter, New York.

L. Devillers, H. Maynard, and P. Paroubek. 2002a. Méthodologies d'évaluation des systèmes de dialogue parlé : réflexions et expériences autour de la compréhension. In *Traitement Automatique des Langues*, volume 43, pages 155–184.

L. Devillers, S. Rosset, H. Bonneau-Maynard, and L. Lamel. 2002b. Annotations for dynamic diagnosis of the dialog state. In *LREC2002*, Spain, May. ELRA.

L. Dybkjær and al. 1998. The disc approach to spoken language systems development and evaluation. In *LREC1998)*, volume 1, pages 185–189, Spain, May. ELRA.

N. Fraser. 1998. *Spoken Language System Assessment*, volume 3. Mouton de Gruyter, New York.

E. Geoffrois, C. Barras, S. Bird, and Z. Wu. 2000. Transcribing with annotation graphs. In *LREC2000*, volume 2, pages 1517–1521, Greece, May. ELRA.

J. Glass, J. Polifroni, S. Seneff, and V. Zue. 2000. *Data collection and performance evaluation of spoken dialogue systems: the MIT experience*.

Lynette Hirschman. 2000. Reading comprehension and question answering new evaluation paradigms for human language technology. In *LREC2000 Workshop "Using Evaluation within HLT Programs: Results and Trends"*, pages 54–59, Greece, May. ELRA.

N. Ide and L. Romary. 2002. Towards multimodal content representation. In *LREC 2002*.

L. Lamel. 1998. Spoken language dialog system development and evaluation at limsi. In *Actes de l'International Symposium on Spoken Dialogue*, Sydney, Australia, November.

MADCOW. 1992. Multi-site data collection for a spoken language corpus. In *DARPA Speech and Natural Language Workshop*.

C. Montacié and G. Chollet. 1997. Systèmes de référence pour l'évaluation d'applications et la caractérisation de bases de données en reconnaissance de la parole. In *16ème JEP*.

R.C. Moore. 1994. Semantic evaluation for spoken-language systems. In *DARPA Speech and Natural Language Workshop*.

P. Sabatier, Ph. Blache, J. Guizol, F. Lévy, A. Nazarenko, and S. N'Guema. 2000. évaluer des systèmes de compréhension de textes. In *Ressources et Evaluation en Ingénierie Linguistique*, pages 265–275. Chibout K. *et al*. (Eds) Duculot.

M. Walker, D. Litman, C. Kamm, and A. Abella. 1998. Evaluating spoken dialogue agents with paradise: 2 cases studies. *Computer Speech and Language*, 3(12):317–347.

M. Walker, R. Passonneau, and J.E. Boland. 2001. Quantitative and qualitative evaluation of darpa communicatorspoken dialog systems. In *Actes du 39me ACL*, pages 515–522, Toulouse, France, July. ACL.

S. Whittaker, L. Terveen, and B. Nardi. 2002. Reference task agenda for HCI. In *ISLE workshop 2002*.