Predictive and objective evaluation of speech understanding: the "challenge" evaluation campaign of the I3 speech workgroup of the French CNRS

Jean-Yves Antoine[1], **Caroline Bousquet-Vernhettes**[2], **Jérôme Goulian**[1], **Mohamed Zakaria Kurdi**[3], **Sophie Rosset**[4], **Nadine Vigouroux**[2], **Jeanne Villaneau**[1]

[1] VALORIA, University of South Brittany, rue Y. Mainguy, F-56000 Vannes, France
Jean-Yves.Antoine@univ-ubs.fr

[2] IRIT, University Paul Sabatier, 118 route de Narbonne, F-31000 Toulouse, France

[3] CLIPS-IMAG, BP 53, F-38041 Grenoble Cedex 9, France

[4] LIMSI-CNRS, Orsay, France

Abstract

This paper presents a new paradigm of "challenge" evaluation of Spoken Language Understanding. This methodology aims at a quantitative assessment with a high diagnostic power, by opposition with standard ATIS-like frameworks. This paper details the methodology as well as the results of an evaluation campaign held by the French CNRS research agency. The benefits of this methodology are also discussed.

1. Introduction

The recent development of spoken language processing has gone along with large-scale evaluation programmes that concern spoken language dialogue systems as well as their components (speech recognition, speech understanding, dialogue management). This paper deals with the evaluation of Spoken Language Understanding (SLU) systems in the general framework of spoken Man-Machine Communication.

1.1. Man-Machine Communication

Man-Machine Communication concerns interactive systems that aim at providing a natural interface between human and computers on task oriented dialogues. It concerns most of the time an information task[1] where the dialog system provides an interface between the user and a database.

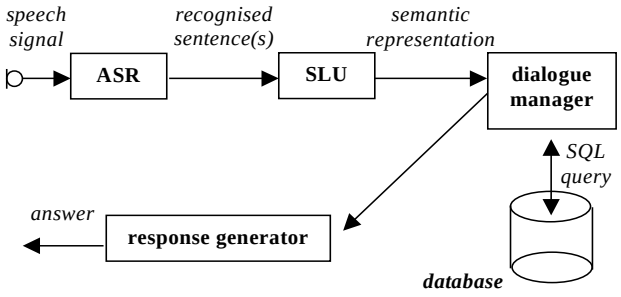

Figure 1 — Generic architecture of a spoken dialog information retrieval system

The general architecture of a spoken information retrieval system is described on figure 1. First of all, a module of automatic speech recognition (ASR) is in charge of providing the system with a sequence or lattice of words from the speech signal (microphone or telephone input). This recognised sentence is supposed to correspond with the utterance that the speaker pronounced. Then, the spoken language understanding (SLU) component builds a semantic representation (for instance, a semantic frame) which stands for the "useful" meaning of the sentence with regard to the task. The dialog manager handles then a contextual analysis to complete this semantic representation[2]. It performs for instance anaphora resolutions. Moreover, it is in charge of the control of the dialog between the user and the computer as well as the generation, when necessary, of a database query. Finally, the answer generation presents (speech synthesis, video display…) the response to the user.

This paper deals only with the evaluation of the spoken language understanding component, that is to say the assessment of its ability to elaborate correctly a semantic structure that should be either contextual or not.

1.2. Evaluation of spoken language understanding

Generally speaking, the evaluation of SLU has always been based on a quantitative metrics that offers an objective and reproducible survey of the system's behaviour. For instance, the DARPA ATIS[3] evaluation programmes follow a *glass box* methodology, where a global accuracy rate is computed through the comparison of the outputs of the system with predefinite references (Hirschman 1998, Dykjbaer *and al.* 1998). Such a

[1] See for instance the well-known ATIS task (Price 1990) where the user can query the system to acquire information about fares and flight schedules.

[2] This bottom-up architecture is relatively caricatured for the purpose of the explanation. For instance, some systems present a integrated architecture where the dialog manager includes the speech understanding component.

[3] ATIS : Airlines Transport Information Systems

quantitative approach provides an interesting survey of the state-of-the-art technology. If a quantitative approach is an essential guarantee of objectivity, it appears consequently that ATIS-like evaluations are often restricted by their global nature. Despite its indisputable interest, it boils down nevertheless to a measurement of the overall performances of the system on a very specific task.

Besides, it could be interesting to be able to compare several SLU systems that concerns different tasks in a diagnostic perspective (behaviour of the systems on complex linguistic phenomena and technological difficulties, speech recognition errors for instance).

Several proposals (Fracas, 1996; Antoine *et al*, 2000) have already been made to achieve such a detailed diagnosis in an objective perspective. These methodologies require however a heavy definition of precise tests suites. This should be why none of these proposals has led to a large-scale evaluation programme.

This paper presents a methodology of "challenge" evaluation that achieves a light but detailed assessment of SLU systems. These systems are concerned with different tasks. Moreover they are based on different approaches. This methodology, which should be considered as a complement of standard ATIS-like methodology, was used during an evaluation campaign founded by the GDR-I3 (*Intelligence-Interaction-Information*) programme of the French CNRS research agency. We first describe this methodology of evaluation. Then we detail the practical achievement and the results of this first evaluation campaign. The benefits and the limitations of this methodology are also discussed. To conclude, we present the objectives of a second campaign of "challenge" evaluation.

2. Methodology of "challenge" evaluation

2.1. Objectives

The principal motivation of this evaluation scheme is to compensate for the two main limitations of standard ATIS-like evaluation paradigms: the lack of genericity and the lack of diagnostic power.

- **Genericity** — The ability of dialogue systems to fit easily the needs of various tasks or applications domains[4] constitutes one of the most important question for the current researches in Man-Machine Communication. The generalisation of the results of ATIS evaluations towards other application domains remains indeed an open issue (Hirschman, 1998b).

- **Diagnostic power** — A global evaluation can only provide a coarse-grained survey that presents a weak diagnostic ability. It is difficult to interpret the overall performances of a system and to distinguish its main sources of errors. This interpretation can nevertheless drive usefully future researches.

This should be achieved by an objective and detailed evaluation scheme that assesses the system on separate collections of tests dedicated to well delimited phenomena. Thus, it makes easier the characterization of the capacities and limitations of each assessed system.

This methodology is founded by four key ideas (Antoine, 2001):

- it is based on an objective evaluation scheme,
- it aims at providing a detailed diagnosis of the behaviour of the system by means of the definition of separate tests sets that are each specific to a precise class of phenomena,
- although this methodology involves a significant number of tests (cf. 3.2.1), it achieves a light evaluation that does not require any adaptation of the systems nor the definition of a common representation scheme.
- this methodology intends to achieve a fruitful comparison of experiences through a common analysis of the error cases of every system.

2.2. Methodology

The methodology of "challenge" evaluation answers the following principles.

A specific tests set for every system — Each system is assessed on a specific set of tests which is elaborated from several *initial utterances*. These initial sentences, which are considered to be representative of the task, are provided by the designer of the system. Since each tests' set is specific to a system, this methodology neither requires the definition of a common task nor common semantic representations. This guarantees a certain lightness of the evaluation.

Challenge — Every tests set is compound of *derived utterances* elaborated from the initial utterances. Each participant proposes a set of derived tests from the initial utterances of the others participants. Each system is thereby assessed on the tests defined by all of the other participants. The derived tests should be considered as a more complicated rewriting of the initial utterances. They are supposed to pose problems to the system: every test challenges the system on a specific phenomenon. For instance, the derived utterance[5] (D), which is supposed to challenge the system on self-repairs, was produced from the initial sentence (I):

(I) *non le matin à six heures environ*
(No on morning around six o'clock)
(D)*non c'est le matin à sept euh non à six heures environ*
(*No it's on morning at seven hum no six o'clock*)

The derivation process is done carefully in order to respect the scope of the task of the assessed system.

Discriminant derivation — In order to achieve a diagnostic evaluation, the derivation process should be as systematic as possible. That means this process should leads to the definition of a collection of tests' sets that are

[4] One distinguishes the concept of field, which corresponds with the universe of realization of the interaction ("railway relation customers "for example) of the concept of task which relates to a specific activity of the field (reservation of ticket for example).

[5] In this paper, the French examples have been literally transcribed in English for explanation purposes.

each specific to a precise class of potential difficulties. For instance, a tests set can assess the recovery of speech recognition errors or the processing of speech disfluencies (see section § 3.2.).

Evaluation — Each system is assessed separately on its own tests' set, the evaluation does not enable a direct comparison between the systems. It allows however a diagnosis based on objective measures. The evaluation answers the following procedure:

- each answer is considered correct or not with regard to the specific needs of the system task,
- the derived tests are divided among several sets that each correspond with a class of phenomena. The evaluation provides therefore several objective error rates that draw a detailed diagnosis of the behaviour of the system.

Synthesis — A synthetic analysis of the behaviour of the system is elaborated from the previous objective measures. This synthesis will be compared with the other participants in the light of the general architecture of the system and of their behaviour on each class of potential difficulties.

3. Evaluation campaign of the GDR- I3

This new methodology of evaluation was used for the first time in a large-scale evaluation campaign organized in the framework of the "Speech Understanding" workgroup[6] of the GDR-I3 research program.

3.1. Presentation of the assessed systems

This evaluation campaign involved four French laboratories: CLIPS-IMAG (Grenoble), IRIT (Toulouse), LIMSI-CNRS (Orsay) and VALORIA (Vannes). The VALORIA laboratory submitted two speech understanding systems (LOGUS and ROMUS) to this evaluation. These systems concerned different domains:

- CLIPS-IMAG (Kurdi 2001): tourist information,
- IRIT (Bousquet-Vernhettes and Vigouroux 2000): railways information (timetable's task),
- LIMSI (Lamel and al. 2000): railways information,
- VALORIA-ROMUS (Goulian and Antoine 2001): tourist information,
- VALORIA-LOGUS (Villaneau, Antoine and Ridoux, 2001): tourist information.

CLIPS-IMAG — The **Oasis** system is based on the Semantic Tree Association Grammar Sm-TAG which is a hybrid formalism combining both syntactic and semantic information in one framework. Oasis is based on a serial architecture compound of six modules that should be divided into three main stages from a functional point of view:

Pre-processing — The pre-processing stage is mainly based on pattern matching techniques and it is intended to correct lexical extragrammaticalities, self-corrections and repetitions.

Parsing — We are using a four step parsing algorithm which is based on the combination of inductive rules to Recursive Transition Networks. The key properties of this algorithm are the use of a partial and selective parsing approach, which allows the system to detect and process the relevant parts of the utterance.

Post-processing — This module is based on semantic meta-rules. It aims at normalising false-starts.

IRIT — The **Cacao** system is based on a conceptual and stochastic approach. The speech understanding process is achieved in two passes. Firstly, the word sequence is decomposed into conceptual segments by a stochastic decoder module. A conceptual segment is a word sequence corresponding with the basic unit of the meaning. The language model of the decoder module is represented by a two-level hidden Markov model. The second pass consists on interpreting the conceptual segmentation to give a semantic representation in term of a key-value pair set.

LIMSI — The LIMSI system is a part of the LIMSI-**ARISE** spoken language dialog system (SLDS) which was developed during the Arise project (den Os *and al*. 1999). The assessed SLU system is composed of two component of this SLDS: the semantic analyser and the contextual understanding component of the dialog manager.

The semantic analysis consists of two steps. At first, the output by the speech recogniser (or the typed sentence) is processed so as to normalise the lexical forms and to use local syntax rules to identify and label some unambiguous concepts. The second step is an analysis carried out by the literal understanding module which provides as an output a semantic frame.

The pre-processing module (semantic analyser) allows a sequence of words to be grouped into a conceptual unit according to local syntax. The syntax is described by rewrite rules. This pre-processing provides a conceptually labelled sentence which can be analysed more efficiently by the literal understanding module.

The literal understanding module generates a semantic frame. The main idea is that this module takes the minimum of decisions so as to avoid misinterpretations in the case of uncertainty.

The ambiguity must then be resolved by the dialog manager according to the dialog context and the task model. Contextual understanding consists therefore of interpreting the utterance in the context of the ongoing dialog, taking into account common sense and task domain knowledge. The semantic frames that result from the literal understanding process are thus reinterpreted using default value rules and qualitative values are transformed into quantitative ones.

VALORIA — The two VALORIA systems follow an approach which is related to some extent to robust natural language parsing. Therefore, they aim at providing deep semantic representations which account for detailed conceptual relations inside the recognised utterance. Although they are answering the same motivations, these two systems are based on different approaches.

The **LOGUS** system is based on a logical approach: the semantic representations of the sentences are logical formula (or conceptual graphs) built by composing λ-terms. In the prototype tested during this first campaign, the analysis was split into two phases: the first was exclusively syntactic and based on the principles of

[6] http://www.univ-ubs.fr/valoria/antoine/gdri3 (in French)

Categorial Grammars. The second was exclusively semantic and was founded on the semantic knowledge related to the objects of the application.

The **ROMUS** system implements SLU in a two stage process that involve NLP techniques of robust parsing. The first stage achieves a finite-state shallow parsing that consists in segmenting the recognised sentence into basic units (chunks adapted to spoken language). The second one, a link grammar parser looks for inter-chunks dependencies in order to build the representation of the semantic structure of the utterance. These dependencies are mainly investigated at a pragmatic level.

3.2. Practical achievement of the evaluation

The evaluation was based on the definition, for every assessed system, of 20 initial utterances proposed by its designer. The other participants produced then 15 derived test sentences. On the whole, each system was assessed on 1200 tests. In order to provide a diagnostic survey of the systems behaviour, the derived utterances were divided among several test suites which each corresponds with a precise class of phenomena.

The choice of the phenomena used during the derivation stage remained on the initiative of each participant. The derived tests reflect thereby the scientific of interests of each participant. Thus, some tests remained close to standard "type A" DARPA-ATIS tests. Some other ones reflect the will to insist on complex linguistic phenomena observed in spoken dialogue corpora. This variety of interests appeared to be very interesting. It led indeed to the consideration of a list of classes of problems which, in our opinion, correspond with essential scientific questions for the years to come. The results of the evaluation campaign were studied according to this typology of potential problems.

- **recovery of speech recognition errors** — Speech recognition errors constitute one of the most insidious and difficult problems for speech understanding. In case of recognition error the semantic analyser works indeed on a sentence which is not the pronounced utterance. Let us consider for instance the intended utterance (A), that is to say a spoken utterance which is supposed to be really pronounced. Whatever the kind of recognition error (insertion, deletion, substitution), two different situations should occur. On the first hand, the recognised sentence is syntactically or semantically incorrect (example A1 below). One should assess here the ability of the understanding component to handle a robust partial analysis of the not altered part of the sentence. On the second hand (example A2), the recognition error preserves an apparent correctness but the meaning of the recognised sentence does not correspond with what the speaker said.

  (A) *vingt novembre* [november, 20th]

  (A1) *veux novembre* [want november]
  (A2) *deux novembre* [november, 2nd]

- **robust modelling of the structural complexity of spoken language** — For evaluation purposes, one should distinguish two different kinds of structural complexity. The first one concerns the existence of multiple-goals requests. For instance, the sample utterance (B1) involve two queries, while the example (B2) integrates a declaration with an information request.

  (B1) *Quel est le premier train partant de Vannes à Paris et quel est le dernier train possible pour le retour le lendemain*
  [What is the first train from Vannes to Paris and what is the last possible return train for the next day]
  (B2*) Mon fils a seulement 4 ans quelles sont les réductions possibles*
  [My son is only 4 years old what are the corresponding reductions]

  One should assess here the ability of the system to detect multiple speech acts inside the sentence as well as its ability to integrate in a query contextual information that should occur in a declarative part of the sentence.
  The second problem concerns the understanding of complex objets with multiple attributes, whose detection should need a deep semantic analysis of the sentence. Such a deep analysis is required in (C1) to detect the semantic relation of place between *turn right* and *post office*.

  (C1) *Vous devrez tourner à droite dans la première rue qui suit la poste.*
  [You will have to turn right on the first corner that follows the central post office]
  (C2*) Ne vous reste-t-il pas de chambre simple ou double avec vue sur la mer.*
  [Don't you have any single or double room that looks out onto the seafront]

  The assessment of structural complexity should concern the correct analysis of coordinations (C2) or modality verbs too.

- **robust processing of speech disfluencies** — Because of the on-line nature of spontaneous speech, speech repairs and other unexpected structures (hesitations, repetitions, repairs, self-corrections, false starts, interpolated phrases...) are very common in spoken dialogues (Heeman and Allen, 1999). These disfluencies break down the regularity of the speakers' utterances, hence unavoidable problems of robustness. The robust processing of speech disfluencies is however essential for the achievement of a natural interaction.

- **robust modelling of word-order variations** — Although they are common in spontaneous speech (Antoine and Goulian, 2001), word-order variations (dislocations, inversions, cleft sentences…) has not been identified for the moment being as a key problem for speech understanding. However, the importance of this question should raise as dialog systems will concern more complex tasks. Then, the identification of the relation between the different concepts of the utterance should require an careful analysis of the syntagmatic order, as shown by the examples below:

(D1) *Quels sont les horaires d'ouverture du Louvre*
[What are the opening hours of the Louvre]
(D2*) Le Louvre ses horaires d'ouverture c'est quoi*
[The Louvre its opening hours what are they]

Moreover, it should be stressed that approaches developed for rigid word-order languages like English or French apply hardly to more variable ones — see for instance (Koo and al. 1995) for Korean.

robust processing of the problems of lexical and/or semantic coverage — This important problem results from the restricted nature of the tasks concerned by Human-Computer dialogue. Indeed, these tasks involve on the whole a vocabulary of about 10000 items, that represents only a very small fragment of any natural language. Since the users are not aware of the restricted size of the system's lexicon, utterances with out of vocabulary words are very frequent. This problem concerns speech recognition first. However, because of the ambiguity of natural languages, it happens frequently that the speaker's intended word, even if well recognised, is used with a different meaning that does not concern the task universe. This word must therefore not be detected as a task concept. Hence some crucial problems of semantic coverage for the speech understanding component.

Table 1— Evaluation campaign of the GDR-I3 : Error distribution rate according each assessed phenomena. A "*" denoted that the % rate is not relevant according to the tasks of the evaluated system.

System	**CLIPS (Oasis)**	**IRIT (Cacao)**	**LIMSI (Arise)**	**VALORIA (Romus)**	**VALORIA (Logus)**
Domain	**Tourist information**	**Railways timetable task**	**Railways information**	**Tourist information**	**Tourist information**
Speech recognition errors	7,0 %	0%	0 %	20%	2%
Structural complexity	12,5 %	6.5%*	0 %	6%	8%
Spontaneous speech disfluencies	9,0 %	6%	18,2 %	17%	32%
Word-order variations	2,3 %	14.9%	9,0 %	6%	3%
Lexical and semantic coverage	69,2 %	72.6%	36,0 %	32%	35%
Others : multiple or specific phenomena	0 %	0 %	36,8 %	19%	20%

4. Results

This paper aims at presenting the methodology of "challenge" evaluation rather than detailing the behaviour of the assessed systems. As a result, we will only linger a few over the individual results that each system obtained.

4.1. Results analysis

We used several objective metrics to quantify separately the behaviour of our systems. As noted previously, a global comparison of the error rates of the systems can not be made directly, as the latter are evaluated on different tests sets. On the contrary, one can draw useful conclusions on the strengths and weaknesses of the systems in the light of their behaviour on every class of tests. In particular, we all used the distribution of the errors of each system according to our classes of problems (see section 3.2) to quantify as much precisely as possible the behaviour of the systems. This objective metrics was used to compare the behaviour of the systems. It should be stressed that it has not been possible to assess all of the systems on the whole typology because some of the tested phenomena overstepped the scope the addressed task. For instance, timetable tasks are marginally concerned by complex structures.

Table 1 presents the distribution of errors of each system according to our typology of problems. We have defined a specific class (*others*) that gathers errors that were corresponding with derived tests involving multiple difficulties but also to some very specific tests. For instance, it appeared interesting to assess some systems on spelling tasks (numbers, timetables…).

In our opinion, this table may be considered as a diagnostic on the weaknesses which must be investigated in priority for every system with regard to their own domain and tasks. We are currently working on a refinement of our typology that will be used in the next evaluation campaigns (see section 5).

4.2. Individual results

The results presented table 1 can be refined. Most of the time, a system encounters indeed difficulties only on one sub-class of problems. We detail in this section some of these subclasses of each of the assessed system.

CLIPS-IMAG (Oasis) We obtained a general recall rate of 92,80 % and a precision rate of 97,32 %. In order to give a detailed insight about the performance of our system, we distinguished four types of causes of errors :

- *Speech recognition errors* — We observed 33 cases of speech recognition errors. In 45,45% of the cases of utterances with speech recognition errors, our system was able to provide a correct analysis.
- *Structural complexity and word order* — We distinguished between 8 different types of complex linguistic phenomena (which corresponded with a total of 717 relevant occurrences). The considered phenomena are: ellipsis, segment insertion, word order change, anaphora, negation, co-ordination, syntactic ambiguity, and relative constructions. The observed average of correct processing of these phenomena is 88,95%. This shows the effectiveness of our approach to process the different forms of complex linguistic structures.
- *Speech disfluencies* — We distinguished between 5 forms of disfluencies: incomplete words, hesitations, repetitions, self-repairs, and false-starts. Our system achieves a total performance of 89,53%. It achieved recalls of 92,30%, 80,95%, and 62,5% respectively on repetitions, self-repairs, and false-starts.
- *Semantic and lexical coverage* — The main reason of lack of lexical and semantic coverage we observed is the lack of data for training our system.

IRIT (Cacao) — The understanding error rate (UER) is calculated by adding-up the number of insertions, deletions and substitutions of key-value pairs. We obtained an global UER of 6.03%. The error category is given in the table 1.

We observe that all derived utterances with speech recognition errors are correctly understood. The percentage of errors for the problems of structural complexity concerns only multiple requests, since the aim of our system was not to deal with this kind of requests. However, the error rate remained low because the test corpus contains fewer multiple requests. All the errors concerning spontaneous speech disfluencies are due to the presence of hesitations inside a conceptual segment. The main cause of errors is due to an insufficient training of the language model (lexical and semantic coverage). Indeed, more than 6% of words in the test corpus are unknown. Unknown words are usually correctly interpreted except when they are indispensable for the interpretation.

LIMSI (Arise) — The understanding error rate (UER) is obtained by adding-up the number of insertions, deletions and substitutions of attribute/value pairs assigned by a modal information. The global UER obtained is of 5%. The results for errors category is given in table 1.

The most important problem results from lexical and semantic coverage problems. Most of these problems are due to errors during the pre-processing stage. There concerned most of the time spontaneous speech disfluencies are due to incise with OOV or OOT words. Word-order variations are badly handled during the literal semantic analysis but the contextual analysis is able to correctly reinterpreted the previous semantic frame. The most part of the last category is concerned by the spelling process and represented 38% of this category and 15% of the total number of errors. Around the half of sentence with spelling cause all spelling errors. Since the system used during this evaluation was not intended to treat spelling process., this result was a good surprise for us. The most part of the errors which occur during the literal analysis are corrected during the contextual analysis.

VALORIA (LOGUS) — Because of the current inachevement of LOGUS, it is difficult to characterise the real weaknesses of our approach. The results of the campaign proved a good robustness of LOGUS on structural complexity, word-order variations and speech disfluencies, with the exception of false starts and interpolated phrases. The analysis of the errors showed that the main cause of this weakness was the absence of syntax during the second phase of the analysis. Half of the errors detected during this evaluation campaign are yet corrected in a second prototype. The latter combines syntax and semantic during the second stage of analysis.

VALORIA (ROMUS) — ROMUS lacks of robustness on recognition errors, in particular when they occur within a chunk. Substitution or deletion of some prepositions due to the recognition process led similarly to significant errors. In dealing with spontaneous spoken disfluencies, the results are promising. Indeed 75% of the cause of these errors concern interpolated phrases within a chunk, whereas hesitations, repetitions and self-corrections are correctly handled. It appears however that repetitions could lead to spurious lexical tagging that perturb the other stage of processing. The decision rate of the lexical tagger that was used during the evaluation has been consequently reduced. Good results are achieved in dealing with word-order variations and complex structures. The chunk parsing is rarely the source of the observed errors and provides reliable cues for semantic disambiguation. However, almost 25% of the errors are due to broken chunks (recognition errors for one half, interpolated phrases for the second half) that are not repeated entirely in the rest of the utterance, what is usually the case in spoken French as far as hesitations or repairs are concerned.

4.3. Results comparison

Several conclusions can be drawn from these results. First of all, one should logically observe that the strengths of a system correspond with the scientific motivations of their designers. For instance, The IRIT system and the LIMSI system were developed during the ARISE project. They are real Spoken Language Dialog Systems that were evaluated within the framework of the project (Baggia *and al*. 1999, den Os *and al*., 1999). They aim at providing robust analyses of real users' queries in a task-oriented dialogue. Thus, more of the development effort was to make the system able to process in a robust way speech recognition errors, OOV (for IRIT) or (for LIMSI) multiple queries. Hence the noticeable results of the systems on theses phenomena.

Besides, the systems developed by the VALORIA aim at processing spoken language in its whole structural complexity. It is thus encouraging to observe that these systems handle correctly complex spoken utterances as well as word-order variations. On the opposite, the complex derived tests that the VALORIA proposed are rather marginal on the tasks studied by the IRIT laboratory. Yet, the Cacao system proved able to behave sometimes correctly on these complex structures for

which it had not been initially conceived. This kind of result, which was not expected by the systems designers, are revealing the interest of the "challenge" methodology. Neither a individual logfiles analysis nor an standard glassbox evaluation scheme can provide such an unexpected diagnosis.

Likewise, the LOGUS and ROMUS systems of the VALORIA present a perfectible robustness when speech recognition errors occur in the utterance. Unlike the others participants, the VALORIA does not develop for the moment being a specific research effort on speech recognition. In the long term, the integration of a state-of-the-art recognition module is however an absolute necessity for the designer of a spoken dialogue system. Since some of the derived tests involved simulated recognition errors, this evaluation was consequently very useful to test the systems on future potential limitations that was not yet considered by their designers.

5. Challenge evaluation : first conclusions

This first campaign of evaluation is an opportunity of drawing conclusions on the benefits and the drawbacks of this new methodology.

5.1. Benefits of the methodology

Unexpectedly, the most significant benefits of the campaign should perhaps be found in the fact that this evaluation favoured scientific exchanges between the participants. This contribution was particularly evident during the derivation stage, which gave rise to a very enriching confrontation between each other's scientific interests. For instance, the laboratories which used to participate in standard evaluation programmes were interested to assess their system on sentences hat were noticeably more complex than those on which they are usually challenged. On the opposite, the participants that used to assess their system mainly on linguistic phenomena — according to a NLP-oriented approach — had to reconsider the influence of input technology (recognition errors) on speech understanding. Thus, the "challenge" methodology provides an opportunity to test the systems on unusual situations that exceed the scope of the task they are assigned to. This should be an interesting benefit in terms of genericity.

In our opinion, this genericity constitutes a significant contribution of the methodology. Indeed, the typology of classes of potential difficulties that we defined is clearly independent of any task and any application domain. Although this observation is obvious, it should be recall from this point of view that this evaluation scheme does not requires the definition of a common task. To some extent, we were indeed able to compare in a unique evaluation campaign the behaviour of systems that were working on different application domains.

In conclusion, this campaign of evaluation provided on the whole a first survey of the strengths and weaknesses of our systems. It enables a brief but interesting comparison of the various approaches followed by the different systems. This is why we have decided to continue this campaign of test in order to reinforce the contribution of the "challenge" evaluation. In a first time, some current limitations of the methodology should however be eliminated.

5.2. Current limitations of the methodology

In its practical realisation, this first campaign presented some insufficiencies that prevented the methodology from expressing all of its diagnostic power. Likewise, the lessons drawn from the comparison of the behaviour of the systems should have been more significant. Nevertheless, we believe that this first experiment confirmed the interests we were expecting from the methodology.

The observed insufficiencies of the methodology concern primarily the process of derivation. Three main problems were picked out:

- **Initial utterances** — The derivation process, and consequently all of the evaluation, appeared to be significantly dependent of the initial sentences. The definition of the initial utterances should therefore be carried out carefully. This is however an unavoidable problem for any evaluation scheme. In the future, our objective will be to define a set of initial sentences that is as representative as possible of the task, in order to avoid any methodological bias.

- **Scope of the task** — During this first experiment, the scope of the tasks of the assessed systems was not delimited precisely. Consequently, some derived tests appeared to be relatively artificial to the designers of the systems. In the future, the lexicon of the application[7] will have to be provided by the designer of the system. This should prevent the other participants from proposing derived tests that exceed the scope of the task.

- **Refinement of the typology** — It appeared (see section 4) that the classes of our typology of problems covered still a too large variety of phenomena. Thus, if this first evaluation succeeded in providing a detailed diagnosis on each individual system, it was much difficult to compare these behaviours. Consequently, it is necessary to detail more this typology. This will favour the realisation of a systematic process of derivation. Our workgroup is currently working on the refinement of the typology.

However, these insufficiencies do not throw the benefits of the "challenge" methodology back into question. They just show that the derivation process must be better controlled. This is precisely the aim of our current work.

6. Conclusion : future works

In the light of this first evaluation campaign, the "challenge" methodology appears to be able to answer the triple objective it was assigned to : objectivity, diagnostic power and genericity. This methodology could certainly be extended in other areas of natural language processing. In particular, it seems that a "challenge" approach should apply easily to Man-Man speech automatic translation as well as to the assessment of multimodal understanding systems.

[7] Or at least a representative toy-lexicon.

However, the immediate objective of our workgroup is not to extend the methodology but on the contrary to reinforce it in the framework of SLU evaluation. In particular, we are currently working on a precise definition of the phenomena involved in our typology of potential difficulties. Our next evaluation campaign will thus concern the specific problem of the processing of interpolated phrases.

7. Acknowledgments

This work was founded by the GDR-I3 of the French CNRS research agency.

8. References

Antoine J.Y. *and al.* 2000. Obtaining predictive results with an objective evaluation of spoken dialogue systems: experiments with the DCR assessment paradigm. Proceedings of the 2nd International Conference on Language Resources and Evaluation, *LREC'2000*, Athens, Greece. [www.univ ubs.fr/valoria/antoine/articles/LREC.ps]

Antoine J.-Y., Goulian J. 2001. Word order variations and spoken man-machine dialogue in French : a corpus analysis on the ATIS domain. Proceedings of *Corpus Linguistics'2001*, Lancaster, UK. 22-29. [www.univ-ubs.fr/valoria/antoine/articles/CL2001.ps]

Baggia P., Kellner A., Pérennou G., Popovici C., Sturm J. and Wessel F. (1999). Language Modelling and Spoken Dialogue Systems – the ARISE experience, Proceedings of the 6th Eurospean Conference on Speech Communication and Technology, *Eurospeech'99*. Budapest, Hungary, Vol. 4, p. 1767-1770.

Bousquet-Vernhettes C., Vigouroux N. 2001. Context Use to Improve the Speech Understanding Processing. Proceedings of *SPECOM'2001*, Moscow. 89-92.

Dybkjaer N.O. *and al*. (1998). The DISC approach to spoken language systems development and evaluation. Proceedings of the 1st International Conference on Language Resources and Evaluation, *LREC'98*, Granada, Spain, vol.1, 185-189.

Goulian J., Antoine J.-Y. 2001. Compréhension automatique de la parole combinant syntaxe locale et sémantique globale pour une CHM portant sur des tâches relativement complexes. Proceedings of *TALN'99*, Tours, France. 203-212. [http://www.univ ubs.fr/valoria/antoine/articles/TALN2001.ps]

Hirschman L. 1998. Evaluating spoken language interaction: experiences from the DARPA spoken language program. *In* Luperfoy S. (Ed.) *Spoken Language Discourse*. MIT Press.

Hirschman L. 1998b. Language understanding evaluations: lessons learned from MUC and ATIS. Proceedings of the 1st International Conference on Language Resources and Evaluation, *LREC'98*, Granada, Spain, 117-122.

Lamel L., Rosset S., Gauvain J.-L., Bennacef S., Garnier-Rizet M., Prouts. B. 2000. The LIMSI ARISE System. *Speech Communication*, 31(4):339-354, Aug 2000.

FRACAS Project. 1996. Using the framework. *Fracas project LRE 62-051*. Deliverable D16. European Community.

Heeman P.A., Allen J. F. 1999. Speech repairs, intonational phrases and discourse markers: modeling speaker's utterances in spoken dialogue. Proceedings of *Computational Linguistics*, 25(4), 527-573.

Koo M.-W. *and al*. 1995. KT-STS: A speech translation system for hotel reservation and a continuous speech recognition system for speech translation, Proceedings of the 4th Eurospean Conference on Speech Communication and Technology, *Eurospeech'95*, Madrid, Spain, pp. 1227-1231.

Kurdi M.Z. 2001. A spoken language understanding approach which combines the parsing robustness with the interpretation deepness. Proceedings of *ICAI'01*, Las Vegas. USA.

den Os E. *and al*. (1999). Overview of the Arise project, Proceedings of the 6th Eurospean Conference on Speech Communication and Technology, *Eurospeech'99*. Budapest, Hungary.

Price P. 1990. Evaluation of spoken language systems : the ATIS domain. Proceedings of the 3rd DARPA Workshop on Speech and Natural Language, Hidden Valley, PA, 91-95.

Villaneau J., Antoine J.-Y., Ridoux O. 2001. Combining syntax et pragmatic knowledge for the understanding of spontaneous spoken utterances. Proceedings of *LACL'01*. In *LNAI 2099*, Springer-Verlag, 279-295. [www.univ-ubs.fr/valoria/antoine/articles/01lacl.ps]