DQR Test Suites for a Qualitative Evaluation of Spoken Dialog Systems : from Speech Understanding to Dialog Strategy

Jean-Yves Antoine, Jérôme Zeiliger, Jean Caelen

DQR Test Suites for a Qualitative Evaluation of Spoken Dialog Systems : from Speech Understanding to Dialog Strategy

Jean-Yves Antoine
VALORIA – Equipage
Université de Bretagne Sud
1, r. de la Loi, 56000 Vannes, France
Email : Jean-Yves.Antoine@univ-ubs.fr

Jérôme Zeiliger
INRS-Telecom
16, pl. du commerce, Ile-des-Soeurs
Verdun, Quebec H3E 1H6, Canada
Email : zeiliger@inrs-telecom.uquebec.ca

Jean Caelen
CLIPS-IMAG
Domaine Universitaire, BP 53
38041 Grenoble Cedex 9, France
Email : Jean.Caelen@imag.fr

Abstract

Generally speaking, the evaluation of spoken dialog systems is based on objective metrics that are supposed to offer a complete survey of the system's behaviour. Unfortunately, this approach boils down to a measurement of the overall performances of the system. Despite its indisputable interest, this quantitative approach lacks some predictive power to enable a really informative evaluation.

On the contrary, we propose a complementary methodology that intends to respect this criterion of predictability. Inspired by some NLP works (Rolbert and Sabatier 1996), it is based on the definition of DQR tests which are specific to each linguistic phenomenon, what warrants the qualitative and predictive nature of the evaluation. This paper describes the adaptation of the DQR methodology to speech dialog systems. We have thus defined a multi-level framework of evaluation that concerns speech understanding as well as the dialog strategy.

This paper focuses mainly on speech understanding. At first, the DQR methodology is reviewed. Then, numerous examples illustrate the practical elaboration of DQR test suites. Finally, we highlight the achievement of an evaluation session on several kinds of real spoken dialog systems.

1. Towards a Predictive Evaluation

Speech recognition is becoming robust enough to be employed shortly in commercial dialog systems. As technology enhancement requires periodic evaluation of the prototypes or systems, spoken dialog evaluation is now an inescapable topic for the speech community. Several paradigms have been defined for the evaluation of speech recognition (Gibbon et al., 1997) as well as dialog strategies (Danieli 1996; Walker 1997).

Many works (Dybkjaer 1995; Vilnat 1996) have shown that a subjective[1] evaluation, based on the user's opinion, is too dependant of the latter to be reliable. As a result, most evaluation paradigms rely on objective et reproducible criteria. Unfortunately, this approach boils down on the whole to a measurement of the overall performance of the system. For instance, in the *glass box* methodology, the evaluation consists in computing an accuracy rate by comparison between the outputs of the system and corresponding predefinite references. This quantitative approach has lead obviously to significant results in the last decade. Despite its indisputable interest, it shows however a limited diagnostic power and is heavily tied to application domains. Still, the evaluation of spoken dialog systems needs more predictability and genericness.

1.1 Predictability

Predictability is certainly the most restrictive lack of the quantitative evaluation of spoken dialog systems. Spoken language is a complex object that involves numerous cognitive processes. Furthermore, the linguistic behaviour of the user of a spoken dialog system depends noticeably on the nature of the application.

Because of this complexity, there is a good chance that a global survey of the overall behaviour of your system will not give really useful information for future improvements:

— How should you interpret the overall performances of a system ? Do they depend on a robust recognition, on a representative model of language, or on the contrary on an effective strategy of dialogue ?

— Likewise, what kind of information should you get from a global accuracy rate when you try to improve a specific component of the system ?

One answer to these problems should be to evaluate separately each component of the system. However, a quantitative evaluation will still present the same lack of predictability. Let us consider, by way of illustration, the understanding component of the dialog system.

— Would an overall rate of accuracy enlighten us about future improvements of this component whereas it can not provide even a rough characterisation of the linguistic phenomena that hold the system in check ?

— Likewise, would such an global evaluation give some information on how the system, or its components, will fit an other kind of application ?

Obviously, such important questions can not be investigated by a quantitative approach. Assessing the overall performance of a system will track improvement over time. But it is not sufficient to guide the development efforts and for instance show some weaknesses of the dialog manager or verify the adequacy of its Knowledge sources. What we need on the opposite is a qualitative evaluation, which examines precisely the behaviour of the system confronted by every phenomenon specific to dialog situations : only a detailed diagnosis of the linguistic and interactive abilities of the system can answer the above questions.

In conclusion, the quantitative approach will always be useful for speech researchers. Nevertheless, the complementary achievement of a qualitative approach will increase undoubtedly the benefits of any evaluation session. Indeed, useful predictive power should be gained this way.

Qualitative evaluation techniques have emerged from some projects of the Natural Language Processing community (Fouvry 1996; FRACAS 1996). This paper investigates the adaptation and extension of these techniques to spoken

[1] Some author (Churcher et al. 1997) employs the term of *qualitative evaluation* to qualify this subjective approach. In this paper, *qualitative* does not refer to this restrictive acceptation.

dialog evaluation. It presents a framework for examining closely the linguistic and interactive abilities of every kind of speech understanding and dialog systems. This evaluation consists of the definition of several test suites that are specific to each linguistic phenomenon. Finally, it is important to note that such a qualitative approach enables furthermore the achievement of a generic evaluation.

1.2 **Genericness**

Most spoken dialog systems are based on theories and applications that differ noticeably from one to another. Flight transport information services systems (Clementino 1993, Aust 1995) involve for instance a speech understanding component which is based on a statistic parser and a rough template filler (Laface 1992). On the opposite, more complex applications such as computer-aided drawing need a sharp understanding which can only be reached through parsing methods that are close to NLP's ones (Antoine 1996a, 1996b). This diversity explains the difficulties which are generally met when carrying out an evaluation session between several different systems. Likewise, it is essential to define a generic framework of evaluation, so that the latter is not restricted to a specific kind of application.

Since it always focuses on specific phenomena, a qualitative evaluation is supposed to examine the abilities of the system in a way independent of the application as well as of the underlying linguistic model of the system. As a result, qualitative evaluation achieves a high degree of genericness.

2. **DQR Methodology**

The Natural Language Processing community has investigated qualitative paradigms of evaluation for a long time. In particular, the FRACAS[2] European project has defined a general framework to assess understanding systems along a set of common linguistic phenomena they are supposed to handle (FRACAS 1994; FRACAS 1996). The system is then evaluated through a black-box methodology based on the definition of several DQR test suites. Each suite is specific to a particular linguistic phenomenon. Any DQR test is made of three items that are respectively called D (Declaration), Q (Question) and R (Reply). D corresponds to a self-informative data the system is supposed to understand. Question Q concerns the declaration and assesses a precise phenomenon in D. Provided this phenomenon is properly handled by the system, the latter will answer correctly. Since Q is always a close question, the correct answer (R) belongs to the following values : *yes* / *no* / *don't know*. Here is an example of DQR test (FRACAS 1996) that concerns anaphoric resolution :

(D1) *Peter is attending a meeting. He is to chair it.*
(Q1) *Is Peter to chair a meeting ?*
(R1) [Yes]

The same methodology[3] has been adopted in a NLP evaluation program under the sponsorship of the AUPELF-UREF French-speaking agency (Rolbert and Sabatier, 1996).

It should be stressed that the DQR methodology is distinguished by a high degree of genericness. On the one hand, it is totally independent of the representations of the system, by opposition with standard quantitative approaches. The practical achievement of the DQR evaluation will be detailed further. For the moment being, it is sufficient to note that the system answers the question Q by means of a mere comparison of the respective representations of D and Q. As a result, it only needs its own internal representations to answer, whatever the formalism adopted is. This makes a noticeable difference with standard evaluation regimes that are based on a comparison with predefinite output representations.

On the other hand, the definition of test suites that are specific to some precise linguistic phenomena is largely independent of the application domain : the preparation of a task specific evaluation corresponds here to the selection of some suites in an exhaustive database of generic DQR tests.

As a result, no domain or applications models will have to be built for the evaluation purpose, and no common formal representations will have to be adopted.

Thanks to this genericness, it seems that the DQR methodology should apply to any language technology. We thus propose to examine how it could fit with speech understanding and spoken dialog evaluation.

3. **Extending the DQR Methodology towards Structural Analysis and Dialog**

Despite its noticeable genericness, the DQR methodology can not apply directly to spoken dialog systems.

On the one hand, the spoken language differs noticeably from the written one. As a result, spoken language processing addresses to itself specific purposes that the evaluation must reflect. The consideration of ungrammatical structures —repairs, self-corrections... — is certainly the most striking example yet of the differences between the evaluation of spoken and written language systems.

On the other hand, spoken language systems involve several methods (statistical models of language, case-frames selective understanding) that have little in common with the written ones. Although the DQR paradigm keeps independent of the representations of the system, we will see later that these technical differences necessitate some practical adaptations of the methodology.

This paper will precisely present an adaptation of the DQR paradigm that should fit the specificities of speech communication.

3.1 **Structural Analysis**

Numerous linguistic studies have discussed the differences between written and spoken languages (Blanche-Benveniste et al, 1990). By evidence, the main specificity of the spoken language comes from its extreme structural variability. Because of its dynamic and uncontrolled nature, spontaneous speech presents indeed a high rate of unexpected constructions — hesitations, repetitions, self-corrections, word-order alterations, comments,

[2] FRACAS : FRAmework for ComputationAl Semantics)

[3] Likewise, the TSNLP project (Fouvry & Balkan 1996) follows a similar qualitative approach when it defines a large variety of test sets that are each specific to a precise linguistic phenomenon (Lehman et al, 1996).

interruption, etc. — which are however still understandable.
This structural specificity is even higher in French than in English, since spoken French appears to be rather a free word-order language whereas written French does not. A study on a large spontaneous spoken French corpus (Antoine 1995:165) showed for instance that the rate of word-order alterations should amount to thirty per cent of the total number of utterances. This coming on top of the other kinds of spontaneous constructions, it is not unusual that half of the spoken utterances presents an unexpected structure (Antoine 1994).
Spoken dialog systems must master this structural diversity which is on the contrary not problematic in NLP. Thus, the ability of a system to overcome those difficulties is a key point of the evaluation in spoken language technologies. This is why we propose to extend the DQR methodology — which is yet rather interested in semantic inference — to the level of structural analysis, whether it is viewed in a syntactic standpoint (parse tree) or in a semantic one (semantic structure).

3.2 **Dialog**

Another specificity of spoken dialog systems is obviously the interactive relation between the user and the computer. Although it is conceivable that NLP researchers attach a higher importance to the accuracy of the system than to its dialogic competence, this last criterion is of the highest importance for spoken language systems. It should not be forgotten indeed that the natural aspect of the oral interaction is usually considered the main argument in favour of spoken Human-machine Communication.
For the time being, the dialogic competencies of a system are evaluated by means of subjective approaches. Our ambition here is also to extend the DQR paradigm in order to propose an objective evaluation of the dialogic level. In particular, we are investigating the questions of the speech acts interpretation and of the speaker's intention recognition.

4. A Multi-layered DQR Methodology for Spoken Dialog Systems Evaluation

It is essential that the evaluation of spoken dialog systems reflect all the previous levels of competence, from structural analysis to dialog strategy. This is why we propose to integrate the DQR methodology in a multi-level evaluation that concerns the linguistics abilities of the system as well as its dialogic competencies.
Seven levels of evaluation have thus been defined, which concern respectively speech understanding (levels 1 to 3) and dialog (levels 4 to 7). These different levels are first described briefly in this section. We will then review into details the three first levels, which relate specifically to speech understanding.

4.1 **Speech Understanding Levels**

The three first levels are strictly limited to speech understanding, what means that any dialogic dimension is deliberately ignored. The purpose is to assess whether the system understands the speaker's utterances. It is inevitable that the system should require in some cases the dialog context to fulfil this understanding. However, what is only considered here is the semantic content of the speech turns, regardless of the characteristics — speech acts, speaker's intention,... — of the corresponding dialog. As a result, every sentence D corresponds basically to a user's request, or eventually to several speech turns that have been beforehand normalised in order to remove any dialogic bias.

Level 1: structural analysis (literal understanding) — This level concerns the literal understanding of sentences that include neither nor anaphora. The main difficulty for the system comes here from the structural variability of the spontaneous utterances.
Practically, the aim of this first level is to verify whether the system builds a correct semantic representation from the sentence, whatever its semantic formalism could be. According to our qualitative approach, the aim is not to assess the complete semantic structure of the sentence, but on the contrary one of its components. Our extension of the DQR paradigm follows this key idea: every question Q concerns the characterisation of a specific predicate-argument relation within the declaration D. This ensures thereby the genericness of the evaluation, since this relation corresponds to a linguistic reality that is fully independent of the adopted semantic representation. Let us consider the following request:

(2) *What's the next flight from Paris to London ?*

Just suppose we want to verify whether the system has understood that the departure airport is Paris. The following DQR test will answer this question:

(D2) *What's the next flight from Paris to London ?*
(Q2) *Flight from Paris ?*
(R2) [Yes]

The question Q focuses strictly the evaluation on the relation between the main theme of the request D — the *flight* — and its departure attribute — *from Paris*. A correct answer means that the system is able to handle such a basic relation.
Table 1 illustrates the genericness of this evaluation. Indeed, the evaluation can proceed with any kind of system (Table 1), whether its semantic representation is based on thematic case-frames (Bonneau et al. 1993), a semantic cases theory (Antoine 1996a), and any other formalism : a mere comparison of the respective semantic representations of D and Q will provide the answer.

Level 2: implicit understanding — This level concerns the resolution of anaphoric or elliptic references. More specifically, the problem is restricted here to literal references, that is to say references that:

a — Do not exceed the limits of the utterance.
b — Can be solved at a syntactic or a semantic level: no pragmatic or dialogic information should be required for the resolution.

Here are two examples that respect the above constraints:

(3) *Select the device and move it on the left.*
(4) *I need two return-tickets to London as well as to Dover*

In the sentence (5), the anaphoric pronoun *it* refers to the antecedent *device*. It is clear that this anaphora is motivated grammatically. Likewise, the example (6) corresponds to an ellipsis — *as well as* refers implicitly to *I need two return-tickets* — that can be solve by way of some syntactic knowledge.

We will not linger over these constructions that have been discussed by the NLP community for a long time. Just consider by way of illustration two DQR tests that apply to the previous examples:

(D3) *Select the device and move it on the left.*
(Q3) *Move the device ?*
(R3) [Yes]

(D4) *I need two return-tickets to London and to Dover*
(Q4) *Two return-tickets to Dover ?*
(R4) [Yes]

The question Q assesses directly the anaphoric resolution since the implicit segment is replaced by its corresponding reference. A relevant resolution in D guarantees a correct answer by means of a comparison between the respective semantic representation of D and Q. Once again, this internal comparison enables a generic evaluation.

Semantic representation	**Request (D)** *What's the next flight from Paris to London ?*	**Question (Q)** *Flight from Paris ?*	**Answer (R)** *YES* if matching
Thematic case-frames	$\begin{bmatrix}\text{Request} = \text{flight}\\ \text{Departure} = \text{Paris}\\ \text{Arrival} = \text{London}\end{bmatrix}$	$\begin{bmatrix}\text{Request} = \text{flight}\\ \text{Departure} = \text{Paris}\\ \text{Arrival} = \text{London}\end{bmatrix}$	YES
Semantic cases		$\begin{bmatrix}\text{Pred} = \text{flight}\\ \text{Source} = \text{Paris}\end{bmatrix}$	YES

Table 1 — Genericness of the DQR evaluation : literal understanding level

Numerous anaphora or ellipses exceed the limits of the sentence. They are considered precisely in the next level.

Level 3: inference — This level concerns the elaboration of the full meaning of a sentence and completes therefore the evaluation of the understanding process. The full meaning — or *real meaning* (Pérennou 1996) — is mainly revealed through the pragmatic and dialogic context of the utterance. This is why most declarations D will consist here on several sentences that may represent either some complex requests or several speech turns. It should not be forgotten however that the dialog dimension of these speech turns is not investigated here.

More precisely, the change from the literal meaning to the full one is likely to require three different kinds of process that should be assessed separately :

Common sense reasoning (level 3.1) — Many cognitive processes should be grouped together in this category. In the light of the current state of the art in automatic language understanding, it is reasonable to restrict our purpose to logical inferences. This problem has been largely discussed by the FRACAS consortium. Here is for instance an example of DQR test that corresponds to a case of logical conservativity (FRACAS 1996:66).

(D5) *Many great tenors are German ?*
(Q5) *Are there great tenors who are German ?*
(R5) [Yes]

By opposition with NLP motivations, it is undoubtedly true that logical reasoning is of minor interest for most dedicated spoken language systems. However such high level inferences should meet an increasing importance when spoken dialog systems apply to more complex tasks.

For the time being, the assessment of the logical reasoning abilities of spoken language systems is hardly applicable. Apart from some specific cases, this evaluation requires indeed a complex understanding of the question Q itself ! Nor, the understanding abilities of most spoken dialog systems remain too weak for this additional task, unlike standard NLP systems. One should hope however that future spoken systems will be able to process such high level inferences.

Pragmatic reasoning (level 3.2) — Since spoken dialog systems are dedicated to a restricted domain of application, pragmatic reasoning is of the highest importance for speech technologies. At first, this level investigates pragmatic reasoning in a restrictive way: pragmatics means here any knowledge that is specific to the task. Other aspects of pragmatics are considered in the next level of evaluation. On the contrary, the following tests restrict precisely themselves to the pragmatic model of the task :

(D6) *I need a return to Dover.*
(Q6) *Ticket to Dover ?*
(R6) [Yes]

(D7) *I need a return-ticket to Dover.*
(Q7) *Return-ticket from Calais to Dover ?*
(R7) [Yes]

In example (6), the system has to understand that the substantive *return* refers to a return-ticket. In test (7), the system succeeds only if it realises that the user calls from Calais (pragmatic context).

We have decided to distinguish this task-based level for practical reasons : one must define a specific test suite for any domain of application. This non-genericness explains why we do not award a large relevance to this level, despite its inescapable use. Actually, it is rather interesting in a practical point of view than in a scientific one: one the one hand, it is very useful for assessing the real appropriateness of a system to its dedicated application, but on the other hand it is of little significance when you consider the general competencies of the system.

Multiple turns inferences (level 3.3) — By definition, dialog systems have to deal with multiple utterances. Many references see therefore the span from their referentials going deeper across the successive utterances or the speaker's turns. The context needed by the anaphoric resolution extends to several sentences or even the whole dialog. Besides, anaphora and ellipses may combine to give highly unbound statements.

This level investigates the inferences that concern several speech turns, regardless of their dialogic features. Thus, these speech turns are normalised in order to remove any dialogic bias. Just consider, by way of illustration, the following dialog :

(8) (U1) *May I have a single-bed room for one night ?*
(S2) *It will cost you 35 pounds. Does it suit you ?*
(U2) *It's perfect. Oh yes I had completely forgotten. I need also a double bed one for my colleagues !*

Then, suppose you want to verify whether the system has recovered the reference of the anaphoric pronoun *one* of (U2). This pronoun refers to *room* (U1). You should then consider the normalised test below:

(D8) *May I have a single-bed room for one night? I need also a double bed one for my colleagues !*
(Q8) *Need a double bed room ?*
(R8) [Yes]

This normalised test focuses on the semantic content of the dialog that is useful for the resolution. The system's utterance (S2) is therefore overshadowed. The situation may be different in other cases :

(9) (U1) *Good evening. I need a room for the night.*
(S2) *What do you like ? A single bedroom or a double one ?*
(U2) *Oh, I'd prefer the first solution.*

Request (D)	**Question (Q)**	**Answer (R)**
You don't have information about the buses?	*Want to know something ?*	Matching speech acts ?
$D = \begin{bmatrix} \text{Speech\_act} = [\text{Request\_info}] \\ \text{Intention} = GD \\ \text{Semantic\_struct} = SD \end{bmatrix}$	$Q = \begin{bmatrix} \text{Speech\_act} = [\text{Request\_info}] \\ \text{Intention} = GQ \\ \text{Semantic\_struct} = SQ \end{bmatrix}$	YES
$GD = [\ldots]$	$GQ = [\ldots]$	
$SD = [\ldots]$	$SQ = [\ldots]$	

Table 2 — Genericness of the DQR evaluation : speech act interpretation

Pragmatic or common sense inferences have to be call in to understand that this first solution corresponds to a single-bed room ! Besides, the reference of this ellipsis is situated in the system's utterance (S2). As a result, the following normalised test should be considered:

(D9) *I need a room for the night.*
A single-bed room or a double one.
I'd prefer the first solution
(Q9) *Prefer a single-bed room ?*
(R9) [Yes]

It is clear that this level does not investigate any new understanding process: the difficulty rests on the ability of the system to handle inferences upon several speech turns.

This multiple turns evaluation level foreshadows the upper levels of dialog evaluation.

4.2 **Dialog**

The next two levels tackle dialog. Any DQR test considers here a couple of speech turns or a full transaction[4]. According to whether these levels are considered both from the machine viewpoint (system input) or from the user one (system output), levels 4 and 5 or levels 6 and 7 will respectively be addressed.

The evaluation of the dialog strategy is beyond the scope of this paper. We will then describe briefly these levels of dialog evaluation. Anyway, one should acknowledge that the limited linguistic and dialogic abilities of present spoken language systems prevent a close implementation[5] of the DQR methodology to a dialog level of evaluation (see further for a discussion).

Level 4 : speech acts interpretation — Here, the matter is to evaluate if a request, a confirmation, an assertion, or a contest have been respectively received as such by the system. This interpretation refers to the illocutory goal of the current speech act — *intention in action* for Searle (1969; 1983). Just consider, by way of illustration, the following test :

(D10) *You don't have information about the buses?*
(Q10) *Want to know something ?*
(R10) [Yes]

There is an important point here and that is that question Q10 does not refers to the task, unlike the previous linguistic levels, but on the contrary to some dialog state. Considered the present state of the art of dialog systems, the elicitation of such an information — which remains usually implicit in the dialog strategy of the system — is obviously not conceivable. This extension of the DQR methodology towards dialog evaluation must therefore be

[4] e.g. a transaction that stands from a goal being uttered to that goal being reached / or withdrawn / or reached and satisfied.

[5] Such an implementation would require at least some heavy adaptation of the system to the evaluation paradigm, or at worst a useless extension of its linguistic and dialog models.

considered in a middle-term perspective : one should imagine indeed that a comparison between the two dialog representations of D and Q gives the answer (Table 2).

Level 5 : user's intention recognition — This level assesses the handling of in-depth goals in the transactions — *preliminary intention* for Searle (1969;1983). Just consider the following test :

(D11) *Hello, I'd like to know which bus will take me at a grocery store*
(Q11) *Want to go to a grocery store ?*
(R11) [Yes]

The satisfaction of the direct request (*want a bus number*), though the secondary goal, is submitted here to the elicitation of the primary goal (*want a grocery store*). The way such a constraint is handled by the system may affect noticeably the dialog efficiency. An internal comparison of the dialog representations of D and Q — and not of their semantic representation ! — should enable a generic evaluation too.

Request (D) *I'd like to know which bus will take me at a grocery store*	**Question (Q)** *Want to go to grocery store ?*	**Answer (R)** Matching intention?
$D = \begin{bmatrix} \text{Speech\_act} = [\text{Request\_info}] \\ \text{Intention} = [\text{grocery store}] \\ \text{Semantic\_struct} = \text{SD} \end{bmatrix}$ $GD = [\ldots]$	$Q = \begin{bmatrix} \text{Speech\_act} = [\text{Request\_info}] \\ \text{Intention} = [\text{grocery\_store} \\ \text{Semantic\_struct} = \text{SQ} \end{bmatrix}$ $GD = [\ldots]$	YES

Table 3 — Genericness of the DQR evaluation : speaker's intention recognition

The last levels of evaluation imply a qualitative jump, since the user's point of view is now investigated : here is the system asked about its own supposed replies. Once again, such an extension should not be envisaged closely despite its indisputable interest : it is indeed beyond the possibilities of current systems to be aware of their position of interlocutor ! As a result, the last two levels represents above all useful guidelines for a future extension of the DQR methodology :

Level 6 : relevance of the system reply — Here, the answer of the system should be evaluated across every user's request.

Level 7 : relevance of the dialog strategy — Here, the answer of the system should be evaluated at the end of every transaction or dialog. The questions should be : has the transaction succeeded ? Was it efficiently conducted ?

Although the extension of the DQR methodology to the evaluation of the dialog strategy meets some difficulties, its practical achievement for speech understanding is conceivable here and now.

5. Practical Achievement

This practical section tackles two main questions. First of all, it examines how you shall construct an adequate set of DQR test suites. Then, it presents the achievement of the corresponding evaluation on real systems.

5.1 Building DQR Tests Suites

This section presents several tests suites that intend to illustrate the various possibilities of the methodology. In particular, we will focus attention on the definition of positive and negative DQR tests in order to reach a useful sharpness of diagnosis. This section focuses on speech understanding evaluation (level 1), and more specifically on the structural aspects of literal understanding, since the other linguistic levels have been largely discussed in (FRACAS 1996).

Simple tests: key information retrieval and sharper understanding —The structural level aims at assessing the correct characterisation of every predicate-argument relation within the user's utterance. As a result, D corresponds here to a single utterance of the user., Q concerns a specific predicate-argument relation of D and R is the correct answer to Q, according to the declaration. Just consider the following DQR test:

(D12) *I need to go to Granada tomorrow morning.*
(Q12) *Go to Granada ?*
(R12) [Yes]

This example assesses the relation between the verbal predicate *to go* and its argument *Granada*. This test should be considered according different points of view, what warrants the genericness of the methodology : Sheffield should be interpreted either as the *arrival* place (thematic case-frame approach) of the planned travel, the *destination* argument of the verb (semantic case theory), its adverbial phrase of place (syntactic parsing)...
It should be stressed that the question Q is extremely simple, by opposition with NLP tests. In the light of the current state of the art in speech understanding, it is not conceivable to ask the system for a sharp understanding of a complex question. On the contrary, a misunderstanding of the question Q would bias unfortunately the evaluation.
The previous example was concerning a key information of the request. Besides, the DQR framework applies easily to a sharper semantic information:

(D13) *Turn on the right after the white buildings with the red shutters*
(Q13) *Red shutters ?*
(R13) [Yes]

The question concerns here a secondary information : the relation between the colour *red* and the substantive *shutters*. Such a sharp understanding can not be ignored when considering some complex domains of application (Antoine 1996a, 1996b). One should thus regret that the standard regimes of evaluation do not reach such a level of detail. This weakness results directly from the lack of genericness of these evaluations. Most of them are indeed dedicated to the evaluation of information retrieval systems, which do not require such a sharp analysis.

Negative tests — The previous examples were corresponding to positive tests (affirmative answer). Positive tests are useful to control the correct behaviour of the system, but can not give many accounting for its failure. Fortunately, the complementary definition of

negative tests will increase the diagnostic power of the evaluation : the latter are indeed employed to detect and explain precisely the insufficiencies of the system. Given a specific phenomenon, the idea is to define a negative test that should correspond to a conceivable error of the system. Just consider again the declaration (D13) and suppose the system has failed the previous test. The negative test below should give us some information on the cause of this error :

(D14) *Turn on the right after the white buildings with the red shutters.*
(Q14) *Red buildings ?*
(R14) [No]

This test checks whether the system meets difficulties to parse or to understand sentences that do not present a flat semantic structure. A affirmative answer shows indeed that the system has linked the adjective *red* with the wrong nominal phrase. Though artificial, this example serves nevertheless to illustrate how interesting is the definition of negative tests. Thanks to this precise exploration of the abilities of the system, the DQR evaluation is really predictive.

Unexpected spontaneous constructions — According to the previous examples, it is clear that the extended DQR paradigm allows the assessment of every linguistic phenomenon that concerns either structural analysis or literal understanding. In particular, the DQR paradigm applies easily to the study of the unexpected constructions that are very common in spontaneous speech. For instance, example (15) evaluates the detection of a self-correction while the example (16) assesses the accurate recovering of the latter:

(D15) *I want to leave tomorrow evening ... no, sorry ... morning.*
(Q15) *Tomorrow evening ?*
(R15) [No]

(D16) *I want to leave tomorrow evening... no sorry ... morning.*
(Q16) *Tomorrow morning ?*
(R16) [Yes]

Likewise, the following examples investigate the understanding of a declaration that present a word-order alteration. Positive and negative tests aim once again at characterising precisely the behaviour of the system :

(D17) *On the right of the circle, draw a red triangle.*
(Q17) *Draw a triangle ?*
(R17) [Yes]

(D18) *On the right of the circle, draw a red triangle.*
(Q18) *Draw a circle ?*
(R18) [No]

(D19) *On the right of the circle, draw a red triangle.*
(Q19) *Draw on the right of the circle ?*
(R19) [Yes]

(D20) *On the right of the circle, draw a red triangle.*
(Q20) *Draw on the right ?*
(R20) [No]

Thus, the DQR methodology, that was chiefly designed for the assessment of high level inferences (level 2 and 3), can be perfectly extended to survey the structural analysis of spontaneous speech.

Dialog evaluation — As mentioned before, a DQR evaluation of the dialog strategy can't be envisaged in the short term. This observation should not be interpreted as a weakness of the DQR methodology, but on the contrary as a direct consequence of the insufficiencies of current dialog systems. For instance, the speech act interpretation of a utterance and the user's intention may be successfully investigated, provided the dialogue module can elicitate its own internal representations. Just consider by way of illustration the next examples, that survey precisely the identification of the user's intention.

(D21) *I'd like to know if you've got a map of Granada*
(Q21) *Want to know if he's got a map ?*
(R21) [No]

The expected answer is negative, since the intention of the user is to have a map of Granada and not merely to know if the tourist office possesses this map. On the contrary :

(D21) *I'd like to know if you've got a map of Granada*
(Q21) *Want a map ?*
(R21) [Yes]

In complex tests, it is clear that the declaration D should describe completely the dialogic context that is necessary to answer.

5.2 Evaluation Sessions

Any DQR evaluation boils down to three key steps (we consider here the understanding level of evaluation) :

1 — Direct comparison of the semantic representations of the declaration D and the question Q : an affirmative answer is provided if the two structures match and a negative one if not.

2 — Accuracy decision : the answer of the system is correct if it corresponds to the expected reply R.

3 — Accuracy rate : a quantitative counting is carried out on each specific phenomenon :

$$\text{Acc} = \frac{\text{number of correct tests}}{\text{total number of concerned tests}}$$

The qualitative nature of the evaluation results from the gathering of these accuracy rate by phenomenon.

Question Q — One issue under discussion here is that the system must understand the question Q to produce an interpretable answer. Now, most spoken dialog systems can not achieve a sharp understanding of complex sentences because of the restricted coverage of their language model. The questions Q must therefore be very simple, in order to avoid any evaluation bias. Practically, the semantic structure of the question Q should be very close to that of the declaration D. More precisely, it corresponds most of the time to a sub-part of the latter, since Q addresses strictly a specific phenomenon in D. This constraint of simplicity applies similarly to the dialog level.

Unification — An other important aspect of the practical achievement of the DQR evaluation is that it does not require a heavy adaptation of the system. Indeed, the system operates with D and Q as it would do with any utterance of the users. Once D and Q have been processed separately, the answer is easily processed through a comparison of their respective representations. At the understanding level, this comparison should be simply based on an off-line unification of the two semantic structures. The table 1 — section 4.1 — shows that the answer is immediate, whatever the adopted semantic

formalisms are. Likewise, the table 2 — section 4.2. — suggests that such a comparison may be extended to the dialog levels too.

6. Conclusion

This paper presents a generic and qualitative methodology for the evaluation of spoken language systems. This paradigm is based on DQR (Declaration-Question-Reply) tests suites that are specific to some precise linguistic or dialogic phenomena. By scrutinising the linguistic and dialogic abilities of the systems at multiple levels, it is intended to provide significant insights for systems improvement. This paper shows that this methodology, that was previously designed for NLP purposes, should be easily extended to spoken language technologies. We thus plan to adopt such a DQR evaluation in a French-speaking assessment program under the sponsorship of AUPELF-UREF (Zeiliger 1997).

Bibliographical References

J.Y. Antoine (1994), Cooperation syntaxe-sémantique pour la compréhension automatique de la parole spontanée, PhD Thesis, INPG, Grenoble, France.

J.Y. Antoine (1995), *Conception de dessin et CHM*, in K. Zreik, J.Caelen, *Le* Communicationnel *pour concevoir*, Europia (Ed.), Paris, France, 161:184.

J.Y. Antoine (1996a), Parsing spontaneous speech without syntax, COLING'96, Copenhagen, Denmark, 47:52.

J.Y. Antoine (1996b), Spontaneous speech and natural language processing — ALPES: a robust semantic-led parser, proc. ICSLP'96, Philadelphia, USA, October 1996.

H. Aust, M. Oerder, F. Seide, V. Steinbiss, (1995), The Philips automatic train timetable information system, Speech Communication, 17 (1995), 249-262.

M. Bates, R. Bobrow, R. Ingria, D. Stallard (1994), The Delphi natural language understanding system, in proc. of the 4th ANLP, Stuggart, FRG, Morgan Kaufman, 132:137.

C. Blanche-Benveniste et al. (1990) Le français parlé, CNRS Editions, Paris, France.

H. Bonneau-Maynard, J.L. Gauvain, L.F. Lamel, J. Polifroni, S. Seneff, (1993), A French version of the MIT-ATIS system, proc. EUROSPEECh'93, 2059:2062, Septemeber 1993.

D. Byron, P. Heeman (1997), Discourse marker use in task-oriented spoken dialog, Proceedings of Eurospeech'97, Vol. 4, p 2223 - 2226.

G.Churcher, E. Atwell, C. Souter. (1997) Generic Template for the evaluation of Dialog Management Systems, Proc. EuroSpeech'97, Vol 4, p 2247-2250.

D. Clementino, L. Fissore, (1993), A Human-machine Dialog System for Speech Access to Train Time Table Information, proc. EUROSPEECH'93, Berlin, FRG, 1863-1866.

M. Danieli, E. Gerbino (1996), Metrics for evaluating dialog strategies in a spoken language system, proc. AAAI Spring Symposium Series : symposium Empirical methods in Discourse Interpretation and Generation.

H. Dybkjeer, L. Dykbjeer, N.O. Bernsen (1995), *Design formalization and evaluation of spoken language dialogue*, proc. of the 9th Twente Workshop on Language Technology, University of Twente, Enschede, Netherlands.

F. Fouvry, L. Balkan, (1996). Test Suites for Quality Evaluation of NLP Products. Proc. of Natural Language processing and Industrial Applications, NLPIA'96, Moncton, New-Brunswick, Canada..

The FRACAS consortium (1994). Harmonising the approaches, in Public Deliverables of the FRACAS Project (A Framework for Computational Semantics), LRE 62-051, Deliverable D7.

The FRACAS consortium, (1996), Using the framework, Public Deliverables of the FRACAS Project (A Framework for Computational Semantics), LRE 62-051, Deliverable D16, January 1996.

D. Gates, A. Lavie, L. Levin, A. Waibel, M. Gavalda, L. Mayfield, M. Woszczyna, P. Zhan (1996), End-to-end evaluation in Janus: a speech-to-speech translation system. Proceedings of ECAI-96 Workshop on Dialog processing in spoken language systems, Budapest, Hungary.

D. Gibbon, R. Moore, R. Winsky (eds), (1997), Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin.

H. Kamp, U. Reyle (1993), From Discourse to Logic, Kluwer, Dordrecht, The Netherlands.

P. Laface, R. De Mori (eds.), (1992), Speech Recognition and Understanding, Springer, Berlin.

S. Lehmann, S.Oepen, S. Regnier-Prost, K. Netter, V. Lux, J. Klein, K. Falkedal, F. Fouvry, D. Estival, E. Dauphin, H. Compagnion, J. Baur, L. Balkan, D. Arnold, (1996), TSNLP - Test suites for Natural Language Processing, Proceedings of COLING-96, Copenhagen, Denmark, 711:717.

E. Levin, R. Pieraccini, (1997), A stochastic model of computer-human interaction for learning dialog strategies, Proceedings of EuroSpeech'97, Vol. 4, p 1883 - 1886.

I. Mel'cuk, (1987), Dependency syntax : theory and practise, Albany, State University of New-York Press.

R. Montague (1973), The proper treatment of quantification in ordinary English, in J. Hintikka (Ed.), Approaches to Natural Language, Reidel, 221:242.

G. Perénnou, (1996), Compréhension du dialog oral. Rôle du lexique dans le décodage conceptuel, proc. séminaire lexique du GDR-PRC CHM, Toulouse, France.

N. Reithinger, M. Klesen (1997), Dialog act classification using language models, Proceedings of Eurospeech'97, Vol.4, p 2235 - 2238.

M. Rolbert, et P. Sabatier, (1996), Evaluation des systèmes de compréhension de textes: Travaux sur l'évaluation des systèmes de traitement automatique du langage naturel : Etude de l'existant. research report, ARC Informatique, Linguistique et Corpus écrits, AUPELF-UREF.

J.R. Searle (1969), Speech Acts, Cambridge University Press, UK.

J.R. Searle (1983), Intentionality, Cambridge University Press, UK.

J. Zeiliger, J.Y. Antoine, J. Caelen, (1996), Vers une méthodologie qualitative d'évaluation des systèmes de compréhension et de dialog oral homme-machine, proc. JST-FRANCIL'97, Avignon, France.