Obtaining predictive results with an objective evaluation of spoken dialogue systems : experiments with the DCR assessment paradigm

Jean-Yves Antoine[1], Jacques Siroux[2], Jean Caelen[3],
Jeanne Villaneau[1], Jérôme Goulian[1], Mohamed Ahafhaf[3]

[1] EQUIPAGE team, VALORIA, Université de Bretagne Sud, IUP Vannes, r. Y. Mainguy, F-56000 Vannes, France.
Email : {Jean-Yves.Antoine, berthele, Jerome.Goulian}@univ-ubs.fr
[2] CORDIAL team, IRISA-LLI, ENSSAT, 6 r. de Kerampont, F-22305 Lannion, France.
Email : siroux@enssat.fr
[3] CLIPS-IMAG, BP 53, F-38041 Grenoble Cedex 9, France.
Email : { Jean.Caelen, Mohamed.Ahafhaf@univ-ubs.fr }@imag.fr

Abstract

The DCR methodology is a framework that proposes a generic and detailed evaluation of spoken dialog systems. We have already detailed (Antoine *et al.*, 1998) the theoretical bases of this paradigm. In this paper, we present some experimental results on spoken language understanding that show the feasibility and the reliability of the DCR evaluation as well as its ability to provide a detailed diagnosis of the system's behaviour. Finally, we highlight the extension of the DCR methodology to dialogue management.

1. Introduction

During the last decade, the development of spoken language technologies has gone along with the achievement of large evaluation programs which concern spoken dialogue systems as well as some of their components (speech recognition, spoken language understanding, dialogue management). Generally speaking, this evaluation is based on the computation of quantitative metrics that intend to offer an objective and reproducible survey of the system's behaviour. For instance, in the *glass box* methodology, the evaluation consists in computing an accuracy rate by means of a comparison between the outputs of the system and some corresponding predefinite references.

Despite its indisputable interest, such a quantitative approach boils down to a measurement of some overall performances which accounts only for the mean behaviour of the system on large corpora representative of a specific application domain. As a result, standard evaluation paradigms present two serious limitations :

- **Predictability** — Such a global evaluation provides only a rough survey that lacks some predictive power to drive future improvements of the system (Polifroni *et al.*, 1998). An evaluation restricted to the overall system's outputs may furthermore present some methodological biases (Minker, 1998).

- **Genericity** — Another limitation of standard evaluation programs results from their lack of genericity. This weakness founds expression in two different ways. On the one hand, the portability of an evaluation program to another application domain remains an open issue (Hirschman, 1998). On the other hand, the definition of common external predefinite references is a painful and time-consuming task that could disadvantage non-standard systems (Antoine & Caelen, 1999).

Within the context of an evaluation program founded by the French-speaking AUF agency, we have proposed a novel paradigm of evaluation (the DCR methodology) to overcome these limitations (Antoine *et al.*, 1998). Inspired by some NLP evaluation programs (Fracas, 1996), the DCR methodology aims at achieving :

- an **objective evaluation** through the definition of quantitative metrics.
- a **generic evaluation**, since it works simply on the own internal representations of each system : no common representation scheme is needed.
- a **predictive evaluation**, by means of the definition of separate tests suites that assess the system on precisely defined phenomena. This specialisation favours moreover the portability of the evaluation from an application domain to another (see section 6.1).
- a **multi-criteria evaluation**, by means of a test characterisation test withseveral properties that acount for some syntactic, semantic, or pragmatic considerations (see section 3). Multiple diagnoses can thereby be obtained from a unique session of evaluation.

The theoretical bases of the DCR methodology were already presented in (Antoine *et al.*, 1998) and (Antoine and Caelen, 1999). In this paper, we detail the practical achievement of the DCR methodology. At first, the methodology is shortly reviewed. We then present into details the test features on which is based every DCR diagnosis. We then review the lesson that we can draw from preliminary experiments on the DCR evaluation of spoken language understanding :

- practical elaboration of DCR test suites,
- investigation of possible methodological biases,
- comparison with standard evaluation paradigms.

We finally highlight the extension of the DCR methodology to the evaluation of dialogue management.

2. Overview of the DCR methodology

This section presents only a brief overview of the DCR methodology. A detailed description of the paradigm can be found in (Antoine *et al.*, 1998 ; Antoine & Caelen, 1999). It can be downloaded at the following URL: http://www-iupva.univ-ubs.fr/public/IUP/recherche/JYA/

2.1. DCR tests

The DCR methodology is based on the definition of large collection of tests suites where every test is dedicated to the assessment of a unique linguistic phenomenon[1]. The assessed phenomenon is characterised by a set of features that are associated with the test definition (see section 3). Considering the evaluation of *speech understanding* (or more precisely *spoken language understanding*[2]), every DCR test consists of three items :

- the **Declaration D** corresponds to an ordinary user's utterance.

- the **Control C** is a supervised sentence which focuses on a precise phenomenon that is present in the declaration D. Note that the control C corresponds simply to a (correct or incorrect) reformulating of the declaration. It is a simplified utterance that could have been pronounced by an ordinary user.

- the **Reference R** is a boolean value which accounts for the coherence of the two previous utterances ("YES" = correct reformulating).

Here is for instance an example of DCR test that concerns a phenomenon of repairs (negative test: <R> = "FALSE") :

(DCR1) :

<D> *What are the departure flights from Lyon no sorry from Athens.*

<C> *What are the flights from Lyon.*

<R> [FALSE]

When required, the declaration and the control are preceded by the same dialogical context (previous speech turns).

2.2. DCR session of understanding evaluation

Considering a peculiar test, the evaluation proceeds as follows :

1. The declaration and the control are provided separately to the system, which builds two corresponding semantic representations. It should be stressed that the corresponding understanding sessions are totally independent. That is to say that there is no contextual influence of the declaration D on the understanding of the control C.

2. These internal representations are then compared by means of a process of unification. This compatibility test provides a boolean result (YES = compatible semantic representations).

3. The evaluation is considered positive if the compatibility result corresponds to the predefinite reference R.

4. Finally, an overall objective score is provided by collecting the results of the whole test suites. This computation should furthermore take into account the features associated to every test in order to provide a multi-criteria diagnosis of the system's behaviour.

3. Multi-criteria evaluation of spoken language understanding

One aim of the DCR methodology is to achieve a predictive diagnosis which does not boils down to a mere computation of some overall accuracy rates. Since every DCR test is dedicated to a specific phenomenon, it is possible to achieve a detailed diagnosis by means of an adequate selection of appropriate tests. For instance, if you want to assess the robustness of a system uniquely on ellipsis resolution, you will only consider test suites that are dedicated to this phenomenon.

Thus, we have defined a multidimensional system of features which intends to characterise precisely every kind of phenomena assessed by DCR tests. For the moment being, this multi-criteria diagnosis has only been investigated for the evaluation of spoken language understanding. The corresponding features are described in the following subsection.

3.1. Test features

Speech understanding features are grouped into six classes that correspond to different motivations, from syntactic to pragmatic considerations. Obviously, some classes are closely related, whereas some other ones are totally independent. Most of the time, a DCR test should therefore be used for multiple purposes.

3.1.1. Context

Speech understanding is usually viewed as a two stage process : 1) literal understanding concerns the elements of the sentence that can be directly understood, and 2) contextual understanding requires on the contrary the consideration of the dialogic context or of the task universe as well. We propose to distinguish these two kinds of context in the DCR methodology. We thereby have defined a "context" feature that can correspond to three different cases (Table 1).

Value	Description
HCTX	*Context-free* — context is useless for the understanding.
DIAL	*Dialogic context* — the understanding of the assessed element requires the consideration of previous utterances of the dialogue (anaphoric resolution for instance).
TASK	*Task context* — the understanding of the assessed element requires the consideration of the task model (reference computation for instance).

Table 1: Description of the "Context" feature.

3.1.2. Type

This feature characterises the type of information that is tested in the declaration D. Table 2 details these

[1] or simply on a precise part of the utterance as well.

[2] More precisely, this definition of DCR tests corresponds indeed to the evaluation of *spoken language understanding*, and not *speech understanding*. That is to say that the input of the system during the evaluation is not the speech signal, but on the contrary the transcription of the corresponding utterance : the understanding component is assessed without the speech recognition stage of the dialogue system. We are at present investigating the question of the extension of the DCR methodology to the evaluation of speech understanding. Afterwards, we will use indiscriminately the terms of speech understanding and spoken language understanding.

different types. This type is viewed in a rather large meaning, that covers syntactic or semantic considerations (OBJ, PTE types for instance), as well as pragmatic motivations (TYP, MOD types for instance).

Value	Description
TYP	*Type of utterance* — test on the recognition of the dialogic type of the utterance (request, confirmation, answer, etc.).
MOD	*Modality* — test on the recognition of the modality used in the utterance (wish, order, etc.). This modality expresses the illocutory force of the corresponding dialog act.
ACT	*Action* — test on the recognition of the action requested by the user to the system. This action should concern the task (inquiry, reservation, registration, etc.) as well as the dialogue (confirmation, repetition, display, etc.). Most of the time, the action is carried by the main verb.
OBJ	*Object* — test on the recognition of the main object of the utterance (hostel, price, etc.).
PTE	*Object feature* — test on the recognition of one feature of the main object (departure time for a flight, for instance).
ARG	*Argument* — test on the recognition of the other significant elements of the declaration.
SSP	*Argument feature* — test on the recognition of one feature of an argument of the declaration.

Table 2 : Description of the "Type" feature

3.1.3. Syntactic complexity

This feature aims at characterising the influence of the structural complexity on spoken language understanding. Most of speech understanding systems do not implement a detailed syntactic parser. As a result, this feature boils down for the moment being to a rough classification of the main cases of syntactic complexity.

Value	Description
SPL	*Simple* — the assessed element is situated in the main clause of the declaration.
SUB	*Subordination* — the assessed element of the declaration is situated in a subordinate clause.
COO	*Co-ordination* — the assessed element of the declaration is situated in a coordination.

Table 3: Description of the "Complexity" feature

3.1.4. Spontaneous speech

This feature aims at characterising the influence of the ungrammatical nature (hesitations, repetitions, repairs, etc.) of the spontaneous speech on spoken language understansing. Several feature values have been defined, which correspond to the main types of ungrammatical spoken structures (Table 4).

Value	Description
NON	*NULL* — the declaration does not present any ungrammatical spoken structure.
HEU	*Hesitation* — the assessed element of the declaration is situated in (or near) a hesitation.
REP	*Repetition* — the assessed element of the declaration is situated in a repetition.
COR	*Self-correction* — the assessed element of the declaration is situated in a repair.
INZ INC	*Interpolation* — The assessed element of the declaration is situated in (or near) an interpolated clause / phrase.
ANT	*Word-order variation* — The assessed element is subject to some word-order variation (anteposition or other kinds of phrase movement for instance).

Table 4: Description of the "Spontaneous speech" feature

3.1.5. Reference

This feature characterises the way the assessed element is referenced in the utterance. We have distinguished five main classes of reference (Table 5).

Value	Description
NON	*NULL* — the test is not concerned by reference computation.
EXD	*Direct explicit reference* — the assessed element is directly referenced by its name. Example : *the Caumartin hostel.*
EXI	*Indirect explicit reference* — the assessed element is referenced in the declaration by a definite expression. However, the system must carry out some reference computations in order to recover the referenced object. This computation should refer to the task model as well as to the dialogic context. Example : *the first hostel in front of the railway station.*
DEI	*Deictic* — the assessed element is referenced by a deictic expression. Example : *this hostel.*
ANA	*Anaphora* — the assessed element is referenced by an anaphoric pronoun. Example : *the first hostel in front of it.*
ELL	*Ellipsis* — the assessed element is referenced by an elliptic expression. Example : *the same one.*

Table 5 : Description of the "Reference " feature

3.1.6. Nature of the reference

This last feature aims at investigating some specific aspects of reference computation that concern the nature of the referenced object :

Value	Description
NON	*NULL* — the test is not concerned by reference computation.
DEF	*Definition* — test on the identification of the definite / indefinite nature of the referenced element.
NBR	*Number* — test on the identification of the singular / plural nature of the referenced element.

Table 6 : "Nature of the Reference " feature

3.2. Test characterisation

The feature classes detailed in the previous section are finally used to define DCR tests. Thus, every test is associated with a features set that characterises the latter according to these six criteria. Let us consider for instance the following test :

(DCR2) :

<D> *What are the flights to Lyon no sorry to Athens.*
<C> *What are the flights for Athens.*
<R> [YES]

This test evaluates the understanding of the "destination" argument of the declaration D, in spite of the presence of a self-correction. No contextual information is needed. As a result, this test is associated with the following features set (Table 7) :

feature	value
Context	<context> = HCTX
Type	<type> = ARG
Syntactic complexity	<syntax> = SPL
Spontaneous speech	<speech> = COR
Type of reference	<tref> = NON
Nature of the reference	<nref> = NON

Table 7 : Example of test characterisation (DCR2)

It is of first importance to note that the features set characterises the whole test, and not only its declaration D. For instance, in (DCR2), the "destination" argument of the declaration is explicitly referenced by its name ("*Athens*"). However, the test feature "Reference" (<tref>) is not filled with a "EXD" value, since the test does not concern the assessment of reference computation.

On the opposite, the following test (DCR3) concerns reference computation[3] :

(DCR3) :

<D> *What are the flights to Athens.*
<C> *What are the flights from Lyon-Satolas to Athens*
<R> [YES]

In this example, the reference to the "departure" argument is totally implicit (complete ellipsis) and should be solved by a consideration of the task model. The test aims precisely at evaluating this reference resolution. As a result, this test is associated with the following features set (Table 8) :

feature	value
Context	<context> = TASK
Type	<type> = ARG
Syntactic complexity	<syntax> = SPL
Spontaneous speech	<speech> = NON
Type of reference	<tref> = ELL
Nature of the reference	<nref> = NON

Table 8 : Example of test characterisation (DCR3)

4. Development of DCR tests (evaluation of spoken langage understanding)

Since it aims at a detailed diagnosis of the system's behaviour, the DCR evaluation requires the definition of test suites which are significantly larger than in standard [D]ARPA-like evaluation campaigns. By comparison, every request of an ATIS test should lead indeed to the definition of a rather large number of DCR tests which intend to assess exhaustively the understanding of every significant part of the sentence.

The adoption of a systematic methodology of tests building facilitates noticeably this unavoidable effort.

4.1. Test development

In order to ease the development of exhaustive test suites, we propose the following strategy :

1. Definition of test utterances like in usual ATIS-like evaluation programs. When required, these sentences should be associated with a dialogic context. These utterances will be caled *primary declarations*.

2. For every primary declaration, definition of multiple control sentences (C) in order to assess the understanding of every significant part of the declaration. These control sentences are also called *primary controls*. Every association of a primary declaration with a primary control — and the corresponding reference (R) — is called therefore a *primary test*.

One difficulty of this stage results from the definition of negative control sentences. That is to say incorrect reformulating of the declaration; the expert must indeed predict what are the conceivable errors of the systems. However, negative tests are very useful to reach a sharpness of diagnosis that gives many accountings for the system's failures (Antoine *et al*, 1998).

3. Generalisation of the primary tests. Every primary declaration is modified in different ways to account for the various linguistic phenomena that should occur in the utterance. These modifications should concern the syntactic complexity of the declaration (<syntax> feature), the expression of the reference (<tref>, <nref> but also <context>) and the influence of spontaneous speech buildings (<speech>) as well. This generalisation stage leads to the definition of multiple declarations that will be

[3] In this example, the task model assume *Lyon-Satolas* airport to be the default departure, that is to say the airport where is situated the system.

associated with the previous primary controls in order to form the final DCR tests.

Finally, a unique primary should lead to the definition of about fifty DCR tests.

4.2. Tests encoding

DCR test suites are coded in a SGML format that enables a flexible and automatic achievement of the evaluation campaign. The encoding standard of DCR tests files is described into details in (Antoine *et al*, 2000). In this standard, every test is defined by :

- an identifier (test number),
- an set of attributes (features set),
- a declaration D, a control C, a reference and, if necessary, a dialogic context (previous utterances).

Here are for instance some examples of DCR tests that have been built from the same primary declaration ("*Which is the way to the Tour-Eiffel?*") :

```
<test no="20_1" ctxt="HCTX" info="ACT" synt="SPL"
    tref="NON" nref="NON" oral="NON" >
    <D>which is the way to the Tour-Eiffel</D>
    <C> which is the way </C>
    <R>TRUE</R>
</test>

<test no="20_2" ctxt="HCTX" info="ACT" synt="SPL"
    tref="NON" nref="NON" oral="NON" >
    <D> which is the way to the Tour-Eiffel </D>
    <C>how much is it</C>
    <R>FALSE</R>
</test>

<test no="20_8" ctxt="HCTX" info="OBJ" synt="SPL"
    tref="NON" nref="NON" oral="REP" >
    <D> which is the way to well to reach the Tour-Eiffel </D>
    <D> which is the way to the Tour-Eiffel </D>
    <R>TRUE</R>
</test>

<test no="20_12" ctxt="HCTX" info="ACT" synt="SPL"
tref="NON" nref="NON" oral="COR" >
    <D>how much is it well no tell me first what is the way to
    the Tour-Eiffel</D>
    <C> how much is it </C>
    <R>FALSE</R>
</test>

<test no="20_14" ctxt="HCTX" info="OBJ" synt="SPL"
tref="NON" nref="NON" oral="COR" >
    <D>which is the way to the Champ-de-Mars well you see to
    the Tour-Eiffel actually</D>
    <C> which is way to the Tour-Eiffel </C>
    <R>TRUE</R>
</test>
```

4.3. Result files

DCR result files are also coded in a SGML format (Antoine *et al*, 2000). In this encoding standard, every test result is defined by :

- the identifier of the corresponding DCR test,
- the features set of the corresponding DCR test,
- the semantic representations of the declaration D and the control C that have been built by the system,
- the result (SUCCESS / ERROR) of the test.

Since the DCR evaluation is based on the comparison of the internal representations of the system, it is not necessary to keep the semantic structures of D and C in the result file. However, we retain them for the purpose of possible logfiles analyses. Likewise, features sets are kept in the result file : thus, various multi-criteria diagnoses should be carried out *a posteriori* from a unique result file (see section 5.2).

5. Applying theDCR evaluation : results on spoken language understanding

This section accounts for an experiment that intended to verify the practical feasibility of the DCR evaluation. This feasibility study aimed at investigating :

- the possible methodological biases that should go together with the DCR evaluation,
- the possibility of establishing a detailed multi-criteria diagnosis from a collection of DCR tests.

This experiment was achieved on a collection of 251 DCR tests that were based (primary declaration) on the PARISCORP corpus (Rosset *et al.*, 1997). The application domain was tourism information. The evaluation concerned a unique understanding system (LAMBDACOMP, developed by the VALORIA laboratory).

5.1. Methodological biases

One interesting peculiarity of the DCR evaluation is its ability to assess various systems without defining any common representation scheme. From this point of view, the association of a control sentence C with the tested utterance D is the key idea, since the internal comparison of their semantic representations is sufficient to evaluate any system.

However, the definition of the control C may involve some biases that must be investigated with care. Considering the comparison of the semantic structures of D and C, any erroneous understanding of the control C would involve indeed an incorrect evaluation : this misunderstanding would hide any incorrect processing of the assessed declaration D, while a correct understanding of the declaration would involve on the contrary a wrong error detection.

Consequently, the reliability of the DCR evaluation relies imperatively on the definition of control sentences that must be correctly understood by any system, in any situation. This implies that control sentences must be defined with high care :

- the control C must be as simple as possible, in order to facilitate its processing by the system,
- the control C must be as close as possible to its declaration D, the assessed element excepted. Possible evaluation biases are therefore restricted to the only part of the control that differs from the declaration.

In our opinion, these recommendations are sufficient to ensure the reliability of the DCR evaluation. This opinion is supported by a logfiles analysis that was carried out on the 251 tests of the experiment. As a matter of fact, we do not detect any bias on 250 test results. The unique problematic test was the following one :

```
<test no="11_3" ctxt="HCTX" info="ACT" synt="SPL"
tref="NON" nref="NON" oral="NON" >
    <D>What are the hostels near the station</D>
    <C> Where are the hostels near the station </C>
    <R>FALSE</R>
</test>
```

In this negative test, the action requested to the system differs between the declaration (display of a list of hostels) and the control (display of the addresses). However, LAMBDACOMP does not distinguish these two kinds of actions. This is why the system gave erroneously to the control C the same semantic representation as the declaration D.

This error should no be interpreted as a failure of the methodology. Indeed, this bias does not result from a the definition of the control, but rather from a wrong appropriateness of the system to the task addressed by the evaluation. In conclusion, this test has shown some limitation of the genericity of the DCR methodology (see section 6.1), but not really a lack of reliability.

5.2. DCR predictive diagnosis

One aim of the DCR methodology is to provide multiple diagnoses, according various criteria, from a unique collection of tests. Since our experimental study was concerning only 251 tests, it is rather hard to give definitive conclusions on this point. However, this experiment provides some interesting results (Table 9, 10, 11) that suggest that the ability of the DCR evaluation to provide an effective predictive diagnosis.

Number of tests	overall error rate
251	10,5 %

Table 9 : DCR experimental evaluation of spoken language understanding : overall error rate of the LAMBDACOMP speech understanding system

Syntactic complexity	error rate
SPL	9,4 %
COO	16,7 %
SUB	n.s.

Table 10 : DCR experimental evaluation of spoken language understanding (LAMBDACOMP system) : diagnosis on the syntactic complexity.

Spontaneous speech	error rate
NON	4,4 %
REP (repetitions)	11,1 %
COR (self-corrections)	66,6 %
others cases	n.s.

Table 11 : DCR experimental evaluation of spoken language understanding (LAMBDACOMP system): diagnosis on the influence of spontaneous speech.

The table 9 shows the general error rate of the system. This score was computed on the whole tests of the experiment. It provides the same kind of overall result as standard ATIS-like evaluations. Since every DCR test is characterised by several features, it is furthermore possible to refine this rough diagnosis.

Let us consider for instance the "syntactic complexity" feature. The previous overall score can be divided according to the different values of this feature (Table 3). Several scores should then be computed (Table 10), that account for the influence of the degree of structural complexity on the behaviour of the system. For instance, this experiment has shown that the LAMBDACOMP system meets some difficulties to process correctly co-ordinations. As a matter of fact, a significative increase of the error rate (from 9,4 % to 16,7%) was observed between the simple utterances (SPL) and those presenting a co-ordination (COO). The results on subordinate clauses were non significative.

Likewise, the table 11 details the error rates according to the different types of spontaneous ungrammatical structures (repetitions and self-corrections). These results illustrate the influence of spontaneous speech on the robustness of the system. In particular, this detailed diagnosis shows clearly that the current version of LAMBDACOMP is not able to process correctly complex self-corrections.

In spite of their simplicity, these examples illustrate the kind of detailed analysis that should be managed with the DCR methodology. Finer diagnoses, concerning for instance several features in parallel, should be achieved, provided the size of the DCR tests database is statistically significant. One important question concerns precisely the influence of the statistical distribution of the tests according to the different features. It will be investigating by future experiments. Anyway, this question concerns current ATIS-like methodologies of evaluation as well : is a significant number of tests sufficient to assure the reliability of the evaluation ?

6. Conclusion

The development of the DCR methodology is founded by the two main objectives of genericity and predictability. To conclude, we would like to investigate to which extent these objectives have been reached. This question concerns spoken language understanding and dialogue management as well.

6.1. Spoken language understanding

With regard to the question of spoken language understanding, the predictive power of the DCR methodology has not to be proved any longer. Thus, the experiment presented in section five has demonstrated the ability of the DCR evaluation to provide a detailed diagnosis that exceeds easily the possibilities of standard evaluation schemes. This methodology will be applied in the short term in the context of a large scale evaluation program[4] founded by the French-speaking AUF agency. We expect this evaluation campaign to highlight once again the predictive power of the DCR evaluation.

On the opposite, the objective of genericity must be restricted in the sight of the previous experiments. It is undoubtedly true that the DCR methodology is completely independent from the representation schemes. Thus, one

[4] "Speech understanding and spoken dialogue" project (ARC-ILOR B2 : "Dialogue Oral").

does not have to care any more about the elaboration of a common representation scheme, like in standard evaluation programs. However, the practical development of DCR tests has shown that :

- the definition of the control C requires that the expert is rather aware of the behaviour of standard systems. As a result, the independence of the DCR evaluation from the systems is not perfect.
- it is not easy to transpose directly a DCR test suites from one task to another. From this point of view, the DCR methology can not be considered independent from the application domain.

This last judgement must however be restrained. Indeed, the diagnosis provided by the DCR evaluation should be generalised to some extent to various application domains. For instance, if your system is not able to process correctly self-corrections in the ATIS domain, there is every chance that it will behave identically when applied to tourism information ! On the opposite, the overall scores provided by standard ATIS-like evaluation programs can under no circumstances be transposed to another application domain (Hirschman, 1998).

As a result, the DCR methodology seems to be a partial answer to the important questions of portability and genericity. However, this conclusion concerns only spoken language understanding, since the use of the DCR methodology to assess dialogue management remains still an open issue.

6.2. Dialogue evaluation : state of the art

The interaction between the user and the system takes on the most varied aspects. The evaluation of dialogue management must account for all of these features, what explains that many difficulties has prevented for the moment being the achievement of a complete evaluation framework.

Thus, several objective metrics have been proposed (Simpson & Fraser, 1993; Cozannet & Siroux, 1994), that assess the efficiency of the dialogue management according to a certain number of aspects (number of speech turns for instance). However, these metrics can not predict to whole behaviour of the system, and can not detect precisely the weaknesses of its dialogue strategy.

Likewise, the evaluation must account for the user point of view, but the integration of the user's opinion in the evaluation involves some difficulties :

- although it should be combined with some objective metrics (Carletta, 1996 ; Walker *et al*, 1997), the opinion of the casual user is subjective, incomplete and lacks reliability. A solution should be to evaluate several systems with the same group of users (Bonneau-Maynard & Devillers, 1998). This approach presents however the same weaknesses (insufficient coverage, lack of predictability) as standard objective metrics.
- casual users meet diffilculties to express an opinion on precise phenomena (for sintance : relevance of the strategy, relevance of the vocabulary used in a response, etc.). Furthermore, they can hardly give their opinion during the dialogue — any interruption of the interaction will undoubtedly introduce a bias — while one should wonder about the reliability of *a posteriori* opinions, when the dialogic context is different.

6.3. DCR evaluation of dialogue management

These difficulties lead us to propose an extension of the DCR evaluation to dialogue management (Antoine *et al.*, 1998). However, this suggestion supposed that the system had the ability to observe and criticize it own behaviour. Considering the current state of the art in man-machine spoken interaction, this approach is obviously too ambitious.

This is why we propose to investigate a novel methodology of evaluation which rests on two key ideas of the DCR approach — 1) the specialization of the DCR tests, 2) the comparison of the behaviour of the system on two different situations D and C — but which requires the judgment of an expert as well. This methodology consists on :

- providing a **dialogue** (D) whose characteristics and interpretation are perfectly clear for the expert. This dialogue must be elaborated step by step with the system,
- modifying the dialogue D in order to define a **control dialogue** (C) that investigates a precise phenomenon in the dialogue D. This modification must concern uniquely the assessed phenomenon.
- providing the **control dialogue** (C) to the system.
- comparing these two resulting processings with a predefinite reference of compatibiliy (R). This comparison is based on the judgment of the expert, and not on an automatic process.

Let us consider for instance the following dialogue, that will stand for (D) :

(D) :

<U> *I need a single ticket to Paris.*
<S> *When do you want to leave ?*
<U> *Well Monday the 2th.*
<S> *Which is the departure time ?*
<U> *10 o'clock.*
<S> *And from where would you like to leave ?*
<U> *From Lannion*
<S> *Well I can propose you the 3072 TGV departure at 10.30 arrival at 15.15.*

This dialogue ensures that the system is able : 1) to process correctly the different utterances, 2) to exhibit the corresponding dialogue strategy, 3) to provide the requested parameters of the action (departure time, arrival, etc.) as well as a final answer (train number).

Then, one should define for instance the following control dialogue (C) that assesses the ability of the system to process correctly an overloaded answer (<departure date> + <departure time>) :

(C) :

<U> *I need a single ticket to Paris.*
<S> *When do you want to leave ?*
<U> *Well Monday the 2th at 10 am.*

Provided the system is able to process the overloaded answer, an utterance like

<S> *And from where would you like to leave ?*

would be considered correct (R) by the expert. On the contrary, the following answer will be considered erroneous :

<S> *What is the departure time ?*

It should be stressed that the control dialogue (C) does not need to be complete. Finally, the dialogue (D) and (C) should be reformulated in various ways, in order to account for different strategies of dialogue. For instance :

(D) :

<U> *I need a single ticket to Paris.*
<S> *For Paris when do you want to leave ?*
<U> *Well Monday the 2th.*
<S> *Monday the 2th which is the departure time ?*
<U> *10 o'clock.*
<S> *10 o'clock from where would you like to leave ?*
<U> *From Lannion*
<S> *Well I can propose you the 3072 TGV departure at 10.30 arrival at 15.15.*

(C) :

<U> *I need a single ticket to Paris.*
<S> *For Paris when do you want to leave ?*
<U> *Well Monday the 2th at 10 am*

In this case, the following answer will be considered correct :

<S> *Monday the 2th at 10 am from where would you like to leave ?*

On the opposite, each of the follwing answers will be considered erroneous :

<S> *Monday the 2th from where would you like to leave ?*
<S> *at 10 am from where would you like to leave ?*
<S> *Monday the 2th which is the departure time ?*

This methodology should be used in order to assess the following phenomena : cooperative answers, unexpected events (overloaded answers, thematic changes, late corrections, etc.). On the opposite, we don't think that the DCR methodology can investigate easily some aspects that concern the relevance of the reactions of the system. This requires indeed an important human expertise. Nevertheless, we intend to develop the DCR methodology at the dialogue level of analysis, in parallel with its use on speech understanding.

References

Antoine J.-Y., Siroux J., Caelen J., 2000, Evaluation DCR des systèmes de compréhension de parole : norme de structuration des fichiers de tests. research report, EQUIPAGE-LN-2000-1, VALORIA, Vannes, France. http://www-iupva.univ-ubs.fr/public/IUP/recherche/JYA/biblio.html

Antoine J.-Y., Zeiliger J., Caelen J.,1998. DQR test suites for a qualitative evaluation of spoken dialog systems : from speech understanding to dialog strategy. *LREC'1998*. Granada, Spain.

Antoine J.-Y., Caelen J., 1999. Pour une évaluation objective, prédictive et générique de la compréhension en CHM orale : le paradigme DCR. *Langues*. 2(2):130-139. June 1999.

Bonneau-Maynard H., Devillers L., 1998. Evaluation of dialog strategies for a tourist information system, *ICSLP'98*.

Carletta J., 1996, Assessing agreement on classification tasks : the Kappa statistic, *Computational Linguistic*, 22(2), 249-255.

Cozannet A., Siroux J., 1994. Strategies for oral dialogue control, *ICSLP'1994*, Yokohama, 963:966.

FRACAS Consortium., 1996. Using the framework. *Fracas project LRE 62-051*. Deliverable D16, chap. 3.

Hirschman L., 1998. Language understanding evaluations : lessons learned from MUC and ATIS. *LREC'98*. Granada, Spain. 117-122.

Minker W., 1998. Evaluation methodologies for interactive speech systems. *proc. LREC'1998*. Granada, Spain. 199-206.

Polifroni J., Seneff S., Glass J., Hazen TJ., 1998. Evaluation methodology for a telephone-based conversational system. *LREC'1998*. Granada, Spain. 43-49.

Rosset S., Lamel L., Bennacef S., Devillers L., Gauvain J-L., 1997., Corpus oral de renseignements touristiques, technical report, AUPELF-UREF ACR- ILOR-B2.

Simpson A., Fraser N., 1993. Black box and Glass box evaluation of the SUNDIAL system, *Eurospeech'1993*, Berlin, 1423:1426.

Wlaker M., Litman C., Kamm A., Abella, 1997. Paradise : a framework for evaluating spoken dialogue agents, ACL'1997, 271:277.