Toward an objective and generic Method for Spoken Language Understanding Systems Evaluation: an extension of the DCR method

Mohamed-Zakaria KURDI[1 and 2]

[1]Natural Interactive Systems Laboratory (NISLab)
University of Southern Denmark
Main campus: Odense University
Science Park 10
DK-5230 Odense M, Denmark
Kurdi@nis.sdu.dk
http://www.nis.sdu.dk/~kurdi/

Mohamed AHAFHAF[2]

[2]Laboratoire CLIPS – IMAG
BP. 53
380401, Grenoble cedex 09, France
mohamed.ahafhaf@imag.fr

Abstract

In this paper, we present an extension of the DCR method, which is a framework for the deep evaluation of Spoken Language Understanding (SLU) Systems. The key point of our contribution is the use of a linguistic typology in order to generate an evaluation corpus that covers a significant number of the linguistic phenomena we want to evaluate our system on. This allows to have more objective and deep evaluation of SLU systems.

1. Introduction

During the last decade, there was an increased interest in spoken language dialogue systems and especially in their Spoken Language Understanding (SLU) components. Many approaches of spoken language with different theoretical backgrounds were proposed and implemented. This necessitated the development of different evaluation methodologies in order to test the effectiveness of these different approaches. The main common methodologies are quantitative ones like the ATIS evaluation campaign in which the performance of the tested system is measured by comparing its real output with a corresponding analysis by hand.

Despite their interest, these methods do not provide a detailed diagnostic of the negative and positive aspects of the system in term of linguistic phenomena processing. Further more, they require a lot of adaptations (precise task, system's output format, etc.) in order to make an objective comparison between different systems.

To avoid the limitations of quantitative methods, several deep schemes were proposed. Among these schemes, the DCR (Declaration, Control, Reference) method seems the most ambitious to provide a general framework for a qualitative evaluation of spoken language systems (Zeiliger et al., 1997), (Antoine et al., 1998). Despite the improvement of the evaluation quality with this method, it lacks of systematicity that makes the comparison of the results of different systems hard to do. In this paper we present an extension of the DCR method that allow to provide both deep and systematic evaluation.

The outline of this paper is as follows: in section two we present the major requirements of an objective evaluatin method of a SLU system. In section three, we present the main aspects of the DCR method. Our method is described in section four. In section five we provide a description of our experiments and results and finally conclusion and perspectives will close the paper.

2. Major requirements from an objective evaluation method of SLU systems

The major requirements of an objective and generic method for evaluating SLU systems are:

- **Task independence:** the method should be applied to different systems whatever are their tasks.
- **Output format independence and analysis level independence:** one of the major problems that face a generic evaluation method is to be able to compare systems with different output formats or to test systems with different analysis level (syntactic parsing or semantic analysis).
- **Predictivity:** the method should provide a detailed diagnosis of the errors of the system. This allows to drive future improvements of the system.
- **Objectivity:** the evaluation corpus should contain representative linguistic phenomena of the language it is designed to process.
- **Flexibility:** partial evaluation should be possible. For example, one should be able to evaluate his system on a specific phenomenon or a small set of phenomena that he consider as particularly interesting for his system.

3. Presentation of the DCR method

The DCR method was proposed as an attempt to satisfy the major part of the requirement presented above. It is based on the generation of derived test sentences on the basis of initial ones extracted from the corpus on which the system is built. The derived corpus contains a set of groups where every group is dedicated to the evaluation of a unique linguistic phenomenon. Every DCR test consists of three components (Antoine et al., 2000):

1. The Declaration **D**: it corresponds to an ordinary utterance that may be uttered by the system's users.
2. The Control **C**: it consists of a modified version of the utterance D usually with a focus on a precise phenomenon that is present in D.

3. The Reference **R**: it consists of a Boolean value which accounts for the coherence of the utterances C and D.

Here is an example of the DCR test:

<D> I want a double room with with Internet uh Internet connection
<C> I want a double room
<R> False

The main problem of this method is that it does not provide a linguistic framework for the derivation of the D utterances (initial utterances) into C utterances (derived utterances). In fact, the derived utterances are generated following quasi-subjective and task dependent criteria without any guaranty of production systematicity. This makes the comparison of the results of two different systems with different application domains very hard to do.

4. Presentation of our method

In order to overcome the systematicity and derivation objectivity problems in the DCR method, we propose an extended version of it that allows to generate the derived utterances following an a priori defined linguistic typology. The key features of our method are presented in the following paragraphs:

4.1. Initial corpus

The initial corpus consists of a set of utterances relevant to the task of the system. These utterances are chosen following two criteria: in one hand, they have to cover the different semantic aspects of the system and in the other hand, they should provide a riche syntactic base for the derivation operations (they should contain different syntactic structures).

4.2. The derivation grammar

The derivation grammar is built on the basis of syntactic typology that has two main resources:

1. **Existing grammars:** the existing classical grammars and linguistic typological descriptions of the language of the system we want to evaluate are valuable source for the creation of the derivation grammar. They are particularly important because they provide a general and almost exhaustive description of the different standard syntactic phenomena.
2. **Existing linguistic resources:** spoken language corpora are analysed in order to extract the occurrences of different forms of the phenomena we want to test. The major motivation of extracting a part of our rules directly from these corpora is to take into consideration the linguistic phenomena of spoken language that are not systematically considered in the classical grammar books and linguistic typological studies (since they are mainly concerned with written language rather than spoken one).

The transformation grammar contains a set of rules divided into subgroups containing each the set of rules specialized in a specific linguistic phenomena. The rules are written with the following format:

1. **Rules –** two rules are given: the rule corresponding to the structure of the element in the initial utterance on which we want to apply the derivation. This rule is given only when the derivation is applied on a complex structure. The second rule concerns the transformation to be applied.
2. **Transformation type –** we distinguished between two types of transformations:
   a. Internal transformations: they consist of a systematic replacement of some elements inside the test units.
   b. External transformations: they consist of making some operations at the global level of the utterance: by deleting some units, changing their position, etc.
3. **Application conditions –** each derivation rule is associated to a set of application conditions. These conditions are intended to make it precise the nature of test unit to which this transformation operation may be applied. This may lead the human generator in one hand to be systematic in applying the transformations to the whole units to which it might be applied and in the other hand that allows to avoid the generation of agrammatical or semantically inconsistent utterances (especially if the generation is done by a non native speaker).

Two examples of derivations rules with their application conditions are presented below:

1. **An example of an internal transformation rule:**

Rule: Sn (sp)[1]→ pas Sn
[NP (PP) → not NP]
Type: Intra-unit derivation.
Application conditions: this rule may be applied to each non-pronominal Sn (NP) in an elliptic context. For example it cannot be applied to the Sn *une chambre* (a room) in a context such: *je voudrais réserver* <u>*une chambre*</u> (I want to reserve <u>a room</u>)[2].
Example:
This rule may be applied to the elliptical utterance: *une chambre* (a room) which becomes after the transformation: *pas une chambre* (not a room).

2. **An example of an external transformation rule:**

Rule: Sn Sp → Sp Sn
[NP PP→ PP NP]
Type: inter-unit derivation.
Application conditions: this rule may be applied to any type of Sn and Sp.
Example: the utterance: *une chambre pour deux personnes* (a room for two persons) becomes after the derivation: *pour deux personnes une chambre*.

4.3. Derived corpus

[1] The elements between brackets are alternatives to the previous ones.

[2] In order to give an idea about the syntactic changes we are giving literal translation of the examples.

The derived corpus is obtained after applying methodologically the transformations operations defined in the derivation grammar to the initial corpus. Contrary to the DCR procedure, the derivation is done by applying a set of predefined transformations on the basic units in the utterance.

4.3.1. Test unit

One of the main weaknesses in the DCR method is that it does not use an objectively predefined method for the segmentation of the input utterance in order to extract the basic units of evaluation. The segmentation of the initial utterance is done following communicative criteria as we proposed for our formalism Sm-TAG (Kurdi, 2001).

Each evaluation unit corresponds to a unique conceptual segment. A conceptual segment is a set (chunk) of words playing a particular semantic/pragmatic role in the utterance. These roles involve a great variety of cognitive and linguistic considerations such that (Androws, 1985):

- **Topicality of the utterance:** in topic comment articulation, some chunks play usually the role of the topic, which indicates what the utterance is about. The comment, which is the remainder of the sentence, provides information about the topic.
- **Given vs. Non-given:** what the system is presumed to know *a priori* (via the task model) vs. what it doesn't know.
- **Importance:** what is forwarded as important vs. what is backwarded as secondary.
- **Specificity:** whether the speaker is referring to a particular instance of an entity or to this entity in itself.

For example, the utterance: *Je voudrais réserver une chambre pour deux personnes* is segmented in the following way with our segmentation criteria: [je voudrais (topic1)] [réserver (comment1)] [une chambre (comment2/topic2)] [pour deux personnes (comment3)]

The main motivation of using these discourse based rather than classical syntactic phrase based units is that this allows us to reduce the number of derivation and to focus mainly on the syntactic transformations that has a significant implication on semantic and pragmatic interpretation of the utterance.

4.3.2. The derivation process

The derivation process consist of transforming the initial utterances into derived ones by mean of the generation rules. As we saw, the generation rules contain a set of general guidelines for the grammar generator in order to avoid overgeneration and other generation problems. The first step in the generation is the segmentation of the initial utterances following the criteria presented in the 4.3.1. Paragraph. The second step consists of applying systematically the whole transformations described in the derivation grammar to the evaluation units that we obtained after the segmentation of the initial utterances. In order to change only one variable at time, each derived utterance consists of the transformed unit plus the rest of the utterance (without any change) except if the derivation described by a specific rule requires the deletion of a part of the utterance. For example, let us take the following initial utterance: <u>*Je voudrais*</u> <u>*réserver*</u> <u>*une chambre*</u> <u>*pour deux personnes*</u>, and the following derivation rule:

verbe → ne verbe pas [verb → pre-negation mark verb post-negation mark]

The previous rule might be applied only to the first unit (since it is the only unit in the utterance with a verbal head). Although the result of the application of the rule is a well formed utterance: *je ne voudrais pas*, the generated utterance is ***je ne voudrais pas*** <u>*réserver une chambre pour deux personnes*</u> since the derivation rule does not require the deletion of any element in the utterance.

In the other hand, if we have a derivation rule such:

Sn Sv Sn → Sn [NP VP NP → NP]

The derived utterance will contain only one Sn (NP) since the deletion of the rest of the elements is a part of the derivation itself.

5. The experiments

5.1. The Oasis system

As a first experiment of our methodology, we choose to make a test of the Oasis system (Kurdi, 2001). This 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. The general architecture of this system is presented in the following figure:

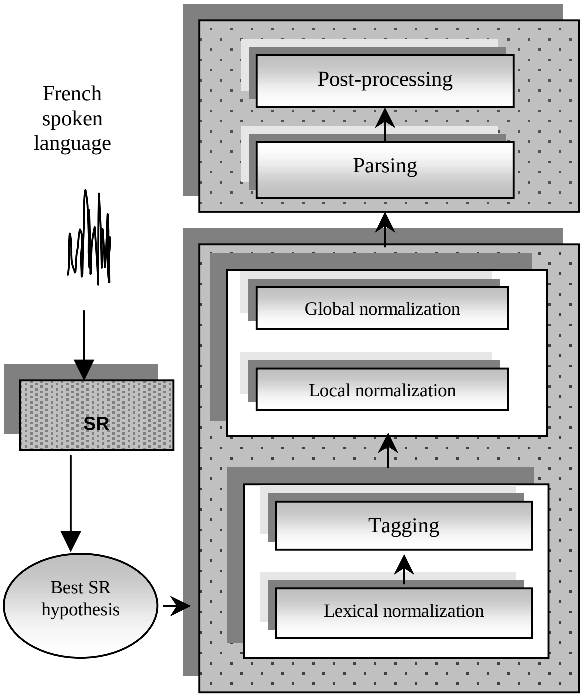

Figure 1. The architecture of our system

As we can see in the previous figure, Oasis system is based on a serial architecture containing 6 modules divided into three main parts from functional point of view:

1. **Pre-processing:** the pre-processing is mainly based on pattern matching techniques and it is intended to correct lexical extragrammaticalities, self-corrections and repetitions.
2. **Parsing:** we are using a 4 step parsing algorithm based on the combination of inductive rules to Recursive Transition Networks RTNs. The key

property of this algorithm is the use of partial and selective parsing approach that allows the system to detect and process the relevant parts of the utterance.

3. Post-processing: we have a post-processing module based on semantic meta-rules intended to normalise the false-starts.

5.2. The considered phenomena

We made an evaluation of this system on three syntactic phenomena that we considered as the particularly relevant for SLU systems. These phenomena are: negation, ellipsis, and extraction.

5.2.1. Negation

Negation is a multidimensional phenomenon that has at the same time lexical, grammatical, and semantic aspects. So, The negation phenomenon is not only a lexical or syntactic reality but also a semantic one. This is one of the main reasons for which we choose the negation as a phenomenon to test our system on. Moreover, in French, there are some interesting differences of negation use between spoken language and written language. For example, the word *ne* (one of the two negation adverbs in French) is often neglected in the informal spoken language like in *je réserve pas* (I reserve not) instead of *je ne reserve pas* in written language and formal spoken language.

We distinguished between three types of negation:

- **Verbal:** when the negation is about a verbal phrase like *je ne voudrais pas une chambre simple* (I do not want a single room).
- **Nominal and prepositional:** it concerns the negations of a nominal or prepositional phrase like: ***pas*** <u>une chambre</u> (not a room), ***pas*** <u>pour une personne</u> (not for one person) (this case is hybrid one: it combines the negation to the ellipsis).
- **Pronominal:** we can have cases like the utterance *rien* (nothing), *aucun* (nobody) (this case is a hybrid one: it combines the negation to the ellipsis).

5.2.2. Ellipsis

The ellipse phenomenon consists of the deletion of one element or more from the utterance without affecting its grammaticality and interpretability. Two major types of ellipsis may be distinguished: grammatical or contextual ellipsis.

- The grammatical ellipsis consists of deleting some words following pure syntactic criteria. For example, in a sentence such *réserves pas* (reserve not) the word *tu* (you) that has the subject function is deleted from the utterance.
- The contextual ellipsis are used frequently in dialogue context in order to avoid the repetition of the already said elements of the utterance. If we consider *je réserve pour demain* at the time of reservation with an agent, this one will understand the request referring to both discourse context and domain of request (ticket, room, etc).

From syntactic point of view, we distinguished between two forms of ellipsis:

- Phrase ellipsis: consist of the deletion of one or more (nominal, verbal or prepositional) phrase from the utterance. For example, *une chambre* (a room) is an elliptical utterance from which the verbal phrase *je voudrais* is deleted.
- Word ellipsis: word ellipsis consists of the deletion of a word playing a specific role in a particular phrase. This word may be the head of the phrase (like the noun in a nominal phrase) or a normal element in it (like a determinant in a nominal phrase). For example, we may have an utterance such *deux* (two), where the noun (which is the head of the phrase) is deleted. In the other hand, we may have an utterance like *chambre simple* (room simple) where the determinant is deleted.

5.2.3. Extraction

The extraction is a phenomenon that allows displacing a phrase (usually prepositional phrase and adverbs) to the right or left of the adjacent phrase without affecting the meaning of the utterance. For example, the adverbial phrase *le 10 décembre à 19 heures 37* (the December 10th at 19 o'clock) in the utterance: *mon train arrive le 10 décembre à 19 heures 37* (my train arrive the December 10th at 19 o'clock) may be displaced to the beginning of the utterance and the transformed utterance becomes: <u>*le 10 décembre à 19 heures 37*</u>*mon train arrive*. The extraction's effect is to divide a sentence into two parts, sometimes on three parts depending on its size and constituents. The extraction is considered as a part of a wide problematic of the words order (Blasco-Dulbecco, 1999) in which we notice the apparition of others phenomenon as double-marking (double-marquage) (Benveniste, 1990) used frequently in spoken language.

We distinguished between different forms of the extraction following the position of the extracted element (preposition or postposition) as well as following the nature of the extracted elements (prepositional phrase, adverb, etc.)

5.3. The generation grammar and derived corpus

We used different grammatical sources in order to write the grammar. These sources include many grammar books like (Gadet, 1989), (Gadet, 1992), and linguistic typological studies like (Benveniste, 1997), (Blasco-Dulbecco, 1999). We also used three spoken language corpora: hotel reservation corpus (Hollard, 1997), Dali project corpus (Sabah, 1997), and Murol corpus (Caelen et al, 1997).

We obtained a total of 154 rules with: 105 negation rules, 17 ellipsis rules, and 32 extraction rules. Some of the rules are hybrid ones (they apply for two phenomena at the same time). These rules cover about 23% of the total number of derivation rules. In order to avoid double generation and allow the independence of the grammar of each phenomenon, the hybrid rules are labelled in a special way in the grammar sets.

In order to limit the number of generated utterances for this first experiment, we generated from one to three utterances corresponding to each rule. The multiple generations were done when we considered that the lexical change might have an effect on the behaviour of the system. Thus, we obtained 252 derived utterances on the basis of ten initial ones.

5.4. Evaluation results

Before we present the results of our evaluations, we resolved two issues:

Selective strategy effect: as we said in a previous section, our parsing algorithm is based on a selective strategy that allows it to detect the relevant part in the utterance. This leaded us to distinguish between two types of generated utterances: relevant utterances and irrelevant utterances. The difference between these two types is that in the relevant utterances the transformation described in the derivation rule is realized in an area relevant for the system (the utterance is then considered as relevant) or irrelevant for the system (the utterance is then considered as irrelevant). Only the relevant utterances were considered in the results calculation.

In the other hand, we considered only the assessed phenomena are considered in our evaluation except if there is an error with the processing of an irrelevant phenomena that was directly caused by a derivation. This limitation allows us to get concentrated only on our targeted phenomena rather than covering the rest.

Following our statistics, 27,8% of the generated utterances was irrelevant to the task of our system. In the other hand, 88,6% of the relevant cases was processed correctly. In only 2,5% of the cases the derivations caused an external error (an analysis error in a non targeted phenomenon). Following our analysis we found that 77,78% of the parsing errors are due to the undergeneration of the grammar while the 22,22% are due to the way in which some rules are implemented.

Below are presented the detailed results sorted by phenomenon.

5.4.1. Negations results

We obtained 157 utterances with negation. The results of the Oasis system on these utterances are presented in the following table:

Type of negation	% of the correctly processed cases
Verbal	91,66
Nominal and prepositional	84,61
Pronominal	78,57
Hybrid with extraction	-
Hybrid with ellipsis	81,59
Total	84,48

Figure 2. Our results on the negation cases

As we can see in the previous table, Oasis system was able to process more easily the classical negation form (the verbal) than the less classical ones, especially the adverbial ones that requires in some cases a higher level of knowledge.

5.4.2. Ellipsis results

Our corpus contains 50 utterances with ellipsis cases. Our evaluation results on these utterances are presented in the following table:

Type of ellipsis	% of the correctly processed cases
Verbal phrase ellipsis	75
Nominal phrase ellipsis	100
Noun ellipsis	0
Determinant ellipsis	71,4
Hybrid: different forms of ellipsis with extraction	100
Total	**76,19**

Figure 3. Presentation of our evaluation results on the ellipsis cases

As we can see in the above table, the Oasis system processing capacity varies following the degree of difficulty of the ellipsis cases. Its capacities are perfect in processing the classical nominal ellipsis cases. Concerning the verbal ellipsis it achieves a coverage of about 75% of the cases. In the case of noun ellipsis, we can see that the Oasis system has a null capacity of processing. This is due to the fact that this kind of ellipsis requires the knowledge of the dialogue context (which beyond the knowledge sources of Oasis) in which this elliptic utterance is realized.

5.4.3. Extractions results

We have 50 utterances with extractions. The results of our evaluation on these cases are presented in the following table:

Type of extraction	% of the correctly processed cases
Preposition	95,45
Postposition	94,54
Verbal	92,72
Nominal and prepositional	96,36
Adverbial	94,44
Total	94,11

Figure 4. Our results on the extraction cases

Our results show that the position of extraction (preposition and postposition) has no real significance for the processing. In the other hand, it shows that the extractions of different constituents are processed in almost the same way although some of them are less frequently observed in spoken language corpora than the rest (like the verbal extractions).

6. Conclusion

In this paper, we presented an extension of the DCR methodology. The main motivations of our extension are:

1. To allow a systematic (and by consequent more objective) generation of the evaluation corpus.
2. To have a more deep diagnostic of the evaluated system.

For satisfying these two conditions, we defined a derivation method that allows to obtain an evaluation corpus build following an a priori defined linguistic typology of the phenomena we want to assess our system on. As we saw, this methodology is task and lexicon independent and allow to evaluate any system independently of the representation level of its output (syntactic, semantic or pragmatic representation).

The application of our method on the evaluation of an SLU system showed that it is realistic and that it allows to obtain a deep diagnostic of the reasons of success and failure of the system.

As a perspective of our work, we intend to apply our method to more than one SLU system (preferably with different approaches) in order to show that it may be used to compare not only the involved systems but also the effectiveness of their approaches to the SLU task.

Finally, we are investigating the possibility of extending our methodology to the evaluation of semantic and pragmatic phenomena in order to enlarge its application domain to the dialogue evaluation.

7. References

ANDREWS, A., (1985), The major functions of the noun phrase, in T. SHOPEN (editor), *Language typology and syntactic description*, Vol. 1 Cambridge university press.

ANTOINE, Jean-Yves, SIROUX, Jacques, CAELEN, Jean, VILLANEAU, Jeanne, GOULIAN, Jerome, AHFHAF, Mohammed, (2000), Obtaining predictive results with an objective evaluation of spoken dialogue systems: experiments with the DCR assessment paradigm, LREC’2000, Athens, Greece.

ANTOINE, Jean-Yves, ZEILIGER, Jérôme, CAELEN, Jean, (1998), DQR Test suites for a qualitative evaluation of spoken dialog systems: from speech understanding to dialog strategy, Proceedings of LREC'98, Granada, Spain.

BENVENISTE, C-B., (1990), *le français parlé : études grammaticales*, Editions du CNRS, Paris.

BENVENISTE, C-B., (1997), *Approches de la langue parlée en français*, Ophrys, Paris.

BLASCO-DULBECCO, M., (1999), *Les dislocations en français contemporain : étude syntaxique*, Honoré Champion, Paris.

CAELEN, J., et al, (1997), Les corpus pour l'évaluation du dialogue homme-machine, ARC B2, Journées JST-FRANCIL, Avignon.

DUBOIS, Jean, et al, (1994), Dictionnaire de linguistique et des sciences du langage, Larousse, Paris.

GADET, F., (1989), *Le français ordinaire*, Paris : Armand Colin, 1989.

GADET, F., (1992), *Le français populaire*, Paris : Armand Colin, 1992.

HOLLARD, Solange, (1997), L’organisation des connaissances dans le dialogue orienté par la tâche, Rapport technique 1-97, GEOD CLIPS-IMAG, Grenoble.

KURDI, Mohamed-Zakaria, (2001), A spoken language understanding approach which combines the parsing robustness with the interpretation deepness, to appear in the proceedings of the International Conference on Artificial Intelligence IC-AI01, Las Vegas, USA, June 25 - 28.

MINKER, W., BENNACEF, S., (1996), Compréhension et évaluation dans le domaine ATIS, Journées d’études de la parole JEP’96, Avignon, France, 417-421.

MULLER, C., (1991), *La négation en français : syntaxe, sémantique et éléments de comparaison avec les autres langues romanes*, Librairie DROZ, Genève.

PICABIA, Lélia, (1975), *Eléments de grammaire générative :* application au français, Paris : Armon Colin.

RIEGEL, M., *et al*, (1994), *Grammaire méthodique du français*, PUF, Paris.

SABAH, Gérard, (1997), Rapport final du projet DALI (Dialogue Adaptatif : Langue et Interaction), http://herakles.imag.fr/pages_html/projets/DALI.html

TESNIERE, L., (1959), *Eléments de syntaxe structurale*, Klincksiek, Paris.

WAGNER, R-L., PINCHON, J., (1991), *Grammaire du français classique et moderne*, Hachette, Paris.

ZEILIGER, Jérôme, CAELEN, Jean, ANTOINE, Jean-Yves, (1997), Vers une méthodologie d'évaluation qualitative des systèmes de compréhension et de dialogue oral homme-machine, actes JST-FRANCIL'97, Avignon, France, 437:446.