Using A Formal Approach to Evaluate Grammars

Bilel Gargouri, Mohamed Jmaiel & Abdelmajid Ben Hamadou

LARIS Laboratory - FSEG - Sfax B.P. 1081 - 3018 SFAX - TUNISIA

Email: Bilel.Gargouri@fsegs.rnu.tn Mohamed.Jmaiel@enis.rnu.tn Abdelmajid.Benhamadou@fsegs.rnu.tn

Abstract

In this paper, we present a methodological formal approach to evaluate grammars based on a unified representation. This approach uses two kinds of criteria. The first one considers a grammar as a resource enabling the representation of particular aspects of a given language. The second is interested in using grammars in the development of lingware. The evaluation criteria are defined in a formal way. In addition, we indicate for every criterion how it would be applied.

1. Introduction

Several research works have been undertaken to propose grammars for natural languages (i.e., English, French, Arab, etc.). For each language, a set of grammars have been suggested depending on its particularities, the aspects to deal with and the choice of an appropriate representing model. Consequently, these grammars are proposed using different formalisms (i.e., Formal Grammars, Unification Grammar, GPSG, HPSG, ATN, etc. [Abeillé, 1997; Miller et al., 1990; Sabah, 1989]). The diversity of these grammars arises a crucial problem of evaluating and choosing the appropriate one needed to handle the aspects of interest. Moreover, the proposed grammars represent a basic language resources for developing a lingware dealing with one or possibly many natural languages. The use of a reliable grammar is strongly required to develop a correct application. In the context of lingware engineering, the grammar evaluation process necessitates an appropriate investigation.

However, few studies have been entirely devoted to the comparison and the evaluation of the different proposals of grammars. These studies are based, generally, on the use of appropriates corpus and routines (i.e., generation, analysis) [Manzi et al., 1996; Sabah, 1989]. They provide, in most cases, some interpretations of these routines application on the relative corpus without recourse to mathematical tools and formal proofs, what limits the credibility of the recommended results. In addition, in the context of lingware development, the evaluation of the proposed grammars has not been considered.

The present framework belongs to a global one in which we deal with natural languages using formal methods in order to implement an environment of formal specification and development of lingware [Gargouri et al., 1998]. In this paper, we deal with the problem of grammar evaluation in two different contexts through the formalisms that describe and represent these grammars. The first context considers a grammar as a resource enabling the representation of particular aspects of a given language. The second one is interested in using grammars in the lingware development. For every evaluation criterion we give a mathematical formula and indicate how it would be applied. In order to realise a rigorous evaluation process for both contexts, we use a unified representation of the main linguistic formalisms that describe the grammars.

We begin with presenting the processing abstraction level of natural languages. Thereafter, we remind the unified (or pivot) representation of grammars. Then, we define the formal tools needed to the grammar evaluation such as sets, functions and predicates. Moreover, we indicate how these tools would be applied. Finally, we present, in a formal way, the evaluation criteria of grammars for the two contexts.

2. Abstraction levels of natural languages

The study of the natural language information system allowed us to observe three abstraction levels of processing. These levels correspond respectively to the natural language the more abstract system, the relative aspects of a language and finally the elementary linguistic knowledge relative to each aspect. The figure below presents these levels:



Figure : Abstraction levels of natural language.

In this description, a natural language (i.e., English, French, Arab, German, etc.) is seen as an abstract system composed of several aspects. The description of one natural language corresponds to the description of all its aspects.

Moreover, an aspect corresponds to a particular phenomena of a natural language (i.e., ellipses, anaphora, etc.). It can concern one or several levels of language processing (i.e., morphologic, syntactic, semantic, etc.). The description of a particular aspect related to one language necessitates the use of an appropriate grammar which must describe the implicated elementary linguistic knowledge. These knowledge (i.e., lexical entry, syntactic category, etc.) represent the lowest level of the present hierarchy.

The evaluation of a grammar can be made w.r.t. natural language, in a general manner, or w.r.t. a particular aspect of a one language. Consequently, the study of the formal evaluation of grammars necessitates the investigation of the relationships between the grammar and the processed aspect as well as between the grammar and the described elementary linguistic knowledge.

3. A unified representation of grammars

It is well known that linguistic formalisms (i.e., Formal Grammars, Unification Grammars, Lexical Functional Grammars, HPSG, ATN, etc.) use a common set of description elements. Indeed, despite the diversity of the used notations and the desired semantics, these formalisms use elementary constituents such as sets, functions, constraints, production rules, maps, features, etc. In order to describe a formalism, one has to specify its elementary constituents and their arrangement way (an appropriate structure to the represented linguistic knowledge).

Starting from these observations, we proposed a pivot representation of the linguistic formalisms that considers all the elementary constituents and the arrangement manners (for each formalism) [Gargouri et al., 1999]. The used unified pivot language is expressive enough since it covers all these formalisms (i.e., a specification language associated to a formal method such as VDM [Cliff, 1986]). The transformations from the original definitions of grammars to the pivot one are realised in a rigorous way in order to preserve their semantics.

The pivot representation of the linguistic knowledge enables to handle them in the same manner. This can facilitate their evaluation and comparison. Moreover, it allows to combine the processing description and consequently to unify the development mechanisms. In this way, we do never depend on linguistic formalisms particularities [Gargouri et al., 1999].

This paper deals with grammars represented in the unified pivot language.

4. Formal tools for the grammar evaluation

In this section, we define the different formal tools needed to formulate the evaluation criteria of grammars such as sets, predicates and functions.

- $L = \{l1, l2, ...\}$: A set of natural languages;
- $A = \{a1, a2, ...\}$: The set of all the aspects of natural languages;

• $C = \{c1, c2, ...\}$: The set of all elementary linguistic knowledge;

- $G = \{g1, g2, \dots\}$: A set of grammars;
- $M = \{m1, m2, \dots\}$: The set of all processing modules;
- *Aspects(l)* : A function that assigns to each language *l* the set of its related aspects;
- $El_Knowledge(a,l)$: A function that assigns to a particular aspect *a* of one language *l* the set of its related elementary knowledge.

• *Decomp(Ling):* A function that associates to a particular lingware *Ling* the set of the elementary aspects to deal with.

• *Modules(Ling)* : A function that gives the set of modules that compose a particular lingware *Ling*.

• Rep(g,a,l): This predicate states that an aspect *a* of a natural language *l* could be represented by a grammar *g*. The implementation of this predicate is based on the use of specific corpus for each aspect. The value to return depends on the possibility of representing the corpus associated to the considered aspect.

• Descr(g,a,l,c): This predicate tells whether an elementary linguistic knowledge related to an aspect *a* of a particular language *l* could be described by a grammar *g*. It can be implemented with a simple procedure that tests whether the language generated by the grammar includes the considered knowledge.

• App(m,g,a,l,c): This predicate informs about the possibility of applying a module *m* on a particular elementary knowledge *c* represented by a grammar *g*, while describing an aspect *a* of a natural language *l*. The implementation of this predicate corresponds to a procedure that verifies this application in the code (or in the description) of the considered module.

• *Coh(c1,c2)*: This predicate verifies the coherence of two elementary linguistic knowledge such as the attribution of a contradictory syntactic categories within a same syntagm of one phrase. To implement this predicate, one can develop a verification procedure for each kind of coherence of interest.

• Redon(c1,c2): This predicate informs on the redundancy of the represented elementary linguistic knowledge such as the description of a same knowledge using different identifiers. It can be implemented with a procedure that verifies if a knowledge has been described needlessly more than once.

• *Mesure_simplicity*(g,a,l): This function measures the simplicity related to the representation of an aspect *a* within a grammar *g*. This function can call others elementary ones that deal with some criteria such as the number of the described elementary constituents, the depth of description, etc.

• $Mesure_efficiency(m, g, a, l)$: This function allows to measure the efficiency associated to the use of one grammar g by a module m while dealing with a particular aspect a of a natural language l. The principle of a such function is well known in the software engineering area [Sedgewick, 1988].

5. The evaluation criteria of grammars

In this section, we describe in a formal way the criteria related to the two contexts proposed above. To each criteria, we associate a formula that enables to evaluate grammars, through the formalisms that describe them w.r.t. this criteria. We deal with these formula at a high level of abstraction.

5.1. General context

In the present section, we consider criteria that allow to evaluate grammars in a general framework of natural language representation. These criteria are the most mentioned in the literature.

• Completness : It allows to evaluate the capacity needed to represent a certain number of expected aspects related to a natural language l. A grammar g meets the property of "completness" w.r.t. a natural language l if it represents all relative aspects to this language.

Completness(g,l) : $\forall a \in Aspects(l)$.*Rep*(g,a,l)=*true*

• *Minimal w.r.t.* Aspect : This criterion allows to ensure that a grammar g represents only the knowledge needed to describe a particular aspect of a language l. A grammar g satisfies the property of "*minimal*" w.r.t. an aspect a of a natural language l if it describes only the needed knowledge.

 $\begin{aligned} & Minimal-A(g,a,l): \ \forall \ c \not\in \\ & El_Knowledge(a,l).Descr(g,a,l,c)=false \end{aligned}$

• *Minimal w.r.t. language* : This criterion states that a grammar g represents only the expected aspects of a language l. A grammar g meets the property of *"minimal"* w.r.t. a natural language l if it don't represent no expected aspects.

Minimal-L(g,l) : $\forall a \notin Aspects(l).Rep(g,a,l)=false$

• *Consistency* : This criterion tells if the linguistic knowledge described by a grammar *g* related to an aspect *a* of a language *l* are not contradictory. A grammar *g* obeys the property of *"consistency"* if all knowledge that it describes are coherent.

 $Consistency(g,a,l): \forall c1,c2 \in C. Descr(g,a,l,c1) \land Descr(g,a,l,c2) \Rightarrow Coh(c1,c2)$

• *No redondancy* : This criterion says if linguistic knowledge described by a grammar *g* related to an aspect *a* of a language *l* are redundant. A grammar *g* has the property of *"no redondancy"* if all knowledge that it describes are not redundant.

 $NoRedond(g,a,l): \forall c1,c2 \in C. Descr(g,a,l,c1) \land Descr(g,a,l,c2) \Rightarrow \begin{tabular}{l} \begin{tabular}{l} Descr(g,a,l,c2) \Rightarrow \end{tabular} \end{tabular} \end{tabular}$

• *Simplicity* : This criterion informs about the representation simplicity of a grammar *g* related to an aspect *a* of a natural language *l*. A grammar meets the property of *"simplicity"* if its measure of simplicity does not exceed a degree *d*.

Simplicity(*g*,*a*,*l*): *mesure_simplicity*(*g*,*a*,*l*)<*d*

• Complementary : This criterion allows to verify if two grammars g1 and g2 are complete each other to represent an aspect a of a natural language l. Two grammars satisfy this property if every elementary knowledge related to the aspect a either is described by g1 or by g2.

• *Multilinguality* : This criterion allows to know if a grammar g allows to represent the same aspect a of two different natural languages *l1* and *l2*. A grammar obeys the property of *"multilinguality"* if it is able to represent the same aspect a of the two languages.

 $Multiling(g,a,l1,l2) : Rep(g,a,l1) \land Rep(g,a,l2)$

5.2. Context of lingware development

In the present context, we are interested in the application of processing, considered as a set of modules, on data or knowledge described by linguistic formalisms.

• Possibility of integration in a lingware : This criterion deals with the possibility of integrating a grammar g, that describes an aspect a of a natural language l, in a lingware Ling. The aspect is implied in the problematic of Ling. A grammar meets the property of "Possibility of integration in a lingware" if one is able to apply modules of the considered lingware on all needed knowledge.

 $\begin{array}{l} PossInteg(Ling,g,a,l): \ \forall \ c \in \\ El_Knowledge(a,l).Descr(g,a,l,c) \Rightarrow \exists \ m \in \\ Modules(Ling).App(m,g,a,l,c)=true \end{array}$

• Adequately to the problem : This criterion allows to know if a grammar g (possibly several) satisfies the requirements of a problem through the representation of a set of particular aspects. Each aspect corresponds to one sub-problem of the whole problem treated by the lingware. A grammar satisfies the property of "Adequately to the problem" if the aspects associated to all sub-problems are represented and integrated in the lingware.

Adequate(p,g,l) : $\forall a \in Decomp(Ling).Rep(a,g,l) \land PossInteg(Ling,g,a,l)$

• *Modularity* : This criterion states if a grammar *g*, in the framework of a lingware *Ling*, can be used partly by the different modules. A grammar *g* meets the property of *"modularity"* if the modules of the lingware are not all applied on the linguistic knowledge related to the processed aspect.

• *Efficiency*: This criterion allows to test the relative efficiency of using a grammar g to represent an aspect a of a language l within a lingware. A grammar satisfies the property "*efficiency*" if the function *mesure_efficiency()* returns a value that is less than a limit e.

Efficiency(*m*,*g*,*a*,*l*) : *mesure_efficiency*(*m*,*g*,*a*,*l*)<*e*

• *Extensibility* : This criterion tells if an extension of knowledge could be brought to a representation of an aspect *a1* in the framework of an already developed lingware *Ling*. This extension allows to augment *a1* in order to obtain a larger aspect *a2* such that $a2=a1\oplus\Delta a$. A grammar meets the property of *"extensibility"* if one can apply modules on all knowledge relative to the resulted aspect.

Extensibility(Ling, g, a1, l) : $\forall c \in El_Knowledge(a2, l)$. $\exists m \in Modules(Ling).App(m, g, a2, l, c)$

6. Conclusion

In this paper, we presented a formal methodological approach for evaluating grammars. The evaluation is based on two sets of criteria. The first one, deals with the grammars performance while describing natural languages. The second is interested in the use of grammars, as a linguistic resources, within lingware.

The present framework belongs to a more large project treated by our team. This project makes use of formal methods to specify and the develop lingware. Our main objective is to establish an environment based on this formal approach. It should be noted that this environment is under implementation. Currently, we are investigating the development of a set of software tools to exploit our formal evaluation approach. These tools may constitute a platform of grammars evaluation.

Finally, we envisage the integration of this platform, as an important component, in our environment. This will permit to make, at earliest stage, the best choice of the appropriate grammars needed to develop a lingware.

7. References

Abeillé Anne and Blache Philippe

Etat de l'art : la syntaxe, TAL, N°38,vol.2, pp69-90,1997. Cliff B. Jones Systematic software development using VDM, Prentice Hall. International 1986. Gargouri B., Jmaiel M. and Ben Hamadou A. Vers l'utilisation des méthodes formelles pour le développement de linguiciels, COLING-ACL'98, 10-14 Août 1998, Montréal, Québec, Canada Gargouri B., Jmaiel M. and Ben Hamadou A. A formal approach to lingware development, IEA/AIE'99, May 31 - June 03, 1999,Cairo, EGYPT(LNCS, Springer-Verlag, n. 1611) Manzi Sandra, King Maghi and Douglas Shona Working towards user-oriented evaluation, NLP+IA 96, Canada, pp.155-160, 1996. Miller Philip et Torris Thérèse

Formalismes syntaxiques pour le traitement automatique du langage naturel, Hermes Paris 1990.

Sabah Gérard

L'intelligence artificielle et le langage : volume 1et 2, Editions Hermès, 1989.

Sedgewick Robert

Algorithms, second edition, Addition Wesley, 1988.