A Knowledge-based Method for Grammatical Knowledge Extraction Process

Jesús Cardeñosa, Carolina Gallardo

Abstract— **This paper describes a new approach for the development of systems that requires natural language parsing or generation. This method is based on the use of Descriptive Grammars –in particular, the descriptive grammar for Spanish is used– as the source for linguistic knowledge extraction. This knowledge source allows the use of classical knowledge-engineering methodologies for the extraction of rules that represent partial or complete aspects of the language, without the necessity of appealing to linguistic theories or experts. This easy method opens a new range of possibilities to the development of reliable applications that require parsing or language generation, dialog systems, information extraction, or semantic web applications.**

***Index Terms*—**Natural Language Processing (NLP); information extraction; language parsing; language generation;

I. Introduction

The dimension of the web permits new uses and new challenges to applications such as knowledge extraction and retrieval, contents dissemination, etc. The web imposes new necessities to technologies that in some way take natural language as input or output, or where multilinguality is a main feature. As such, apart from the enhancement of extraction and retrieval techniques, it is necessary to design components and systems that are able to interact, understand and produce natural language, deriving in the area of natural language processing. The main components in Natural Language Processing (NLP) are analyzers (like parsers of NL queries, PoS (Part of Speech) taggers, analysis modules of machine translation systems, etc.) and language generators.

Most NL-related systems (be it an analyzer or generator) are supported by a natural language grammar. However, the design and development of a grammar to support general domain and robust NLP and information systems is a complicate and expensive task, with high costs in the tasks of design, acquisition of the grammatical knowledge and subsequent expansion, leaving aside the issue of integrating into a single system linguistic knowledge with non-linguistic knowledge.

In this article, we focus on the process of acquisition of linguistic knowledge for grammars supporting NLP systems. More concretely, we are going to describe the followed methodology, which supposes a shift from traditional NLP methodologies for grammar design into Knowledge Engineering methodologies, rarely used in the NLP field.

II. Related Work

Linguistic theories provide a coherent framework for guiding the creation and development of supporting grammars for language analyzers and generators. Issues like the most adequate formalism to express linguistic knowledge, the ontology of a language and the architecture of the grammar are defined by linguistic theories aiming at the construction of grammatical artefacts in computable environments. Further, a linguistic theory pursues coherence and completeness in their account of natural languages and searches for mathematical rigour. In their most applied part, these theories have inspired the so called Grammar Development Environments, which are a set of tools aiding the processes of constructing large scale grammars, together with analyzers and generators based on linguistic theories. Some concrete examples are: PARGRAM [1] based on Lexical Functional Grammar [1] or MATRIX [3] based on HPSG [4]. These grammars have achieved linguistic coverage of about a 55% on free language with a period of development spanning to 5 to 10 years. It could be said that although linguistic theories help, but they are not sufficient in the process of creating computable grammars for natural languages possibly due to insufficient linguistic coverage and other problems.

The process of constructing a grammar supporting real applications (like machine translation, information extraction systems, etc.) comprises a set of techniques whose emphasis lies on the practical aspects of the grammar together as the results. Apart form the theoretical foundations of a system, other topics such as the design of computable grammars to support multilingual systems, the search for a linguistic independent formalism where to codify the grammar or the exploration of other sources of linguistic data and new ways to perform the knowledge acquisition process are sought, such as the knowledge contained in

This work was partially supported by the PATRILEX Project (HUM2005-07260/FILO) of the Spanish Council of Research.

Jesús Cardeñosa is with Group of Validation and Business Applications of the Universidad Politécnica de Madrid; 28660 Madrid, SPAIN. (telephone 0034913521692; e-mail: carde@opera.dia.fi.upm.es).

Carolina Gallardo is with Group of Validation and Business Applications of the Universidad Politécnica de Madrid; 28660 Madrid, SPAIN. (telephone 0034913521692; e-mail: carolina@opera.dia.fi.upm.es)

descriptive grammars [5]. Remarkably, it is in this much more computation-oriented field where attention is directed towards descriptive grammars, the less computable and formalized treaties of language, and thus usually banished in the area of modern linguistics, computational linguistics and natural language processing fields.

This turning-back to descriptive grammars can be due to several reasons (from the linguistic point of view), namely:

A wider coverage of the language descriptions: recent works such as Gramática Descriptiva de la Lengua Española –GDLE hereafter– (*Descriptive Grammar of the Spanish Language*) [6] or the Cambridge Grammar of the English Language [7] (CGEL). GDLE and CGEL constitute complete and exhaustive studies of natural languages (a degree of exhaustiveness and completeness not present in linguistic theories).

- Independence from theories: Descriptive grammars offer linguistic data independently of a given formalism or theoretical postulate which can or cannot be accepted.
- Descriptions show a wider degree of permanency if compared to the permanency of linguistic analysis (limited to the permanency of the theory)

These aforementioned grammars incorporate more structured and comprehensive descriptions of the language, that have been possible to the advances in theoretical linguistic. Meanwhile, the data gathered in a descriptive grammar constitutes new and challenging inputs to linguistic theories: therefore, the linguistic theory and the descriptive grammar grow *a la par*.

This intricate mixture of theory and description produces a very special conjuncture regarding the type of knowledge included in a descriptive grammar. Thus, we can say that descriptive grammars are composed of heterogeneous knowledge: formal and informal, deep and shallow. The formal and deep knowledge of these grammars is derived from linguistic theories, overtly present; whereas the informal and shallow knowledge is constituted by the linguistic descriptions themselves. From the knowledge point of view, they can be described as:

- They are the product of observation of the language use.
- They exist for many languages. Minority languages may be provided with a descriptive grammar, but not for sophisticate linguistic analysis.
- There is not any need for prior training to understand the statements of a descriptive grammar.
- The knowledge of a descriptive grammar is cumulative.

That is, descriptive grammars are a compendium of heterogeneous, non formalized, knowledge about a language, similar to the problems and domains modeled by Knowledge Based systems, where different types of knowledge –deep and shallow [8]- converge and it is offered a framework to computationally treat such type of knowledge. Further, descriptive grammars are so far the most complete source of knowledge of a language. However, their descriptions are still far from being useful (or as initial input) to systems.

In spite of the possibilities that knowledge engineering can offer to the design of NLP systems, both disciplines has rarely converged. Apart from representational issues like the use of frames and feature structures [9], neither the process of grammar design nor the process of acquisition has been aided by knowledge engineering techniques. Some examples of the inclusion of Knowledge Engineering into NLP are the blackboard architecture of Hearsay-II voice recognition system [10], a generation system in the medicine domain [11], or the object-oriented parser of ParseTalk [12].

The next section depicts the use of a descriptive grammar (more concretely, the *Gramática Descriptiva de la Lengua Española*) to develop a knowledge based system analysis and generation system supported by a reversible grammar, with the properties of scalability, maintainability and with a distributed architecture (blackboard), rarely applied to similar applications.

III. The Grammatical Knowledge Elicitation Process

In this section we will describe the knowledge source and the followed methodology for the knowledge elicitation process.

A. The knowledge source: GDLE

GDLE is a descriptive grammar that is by far the most complete and exhaustive work of Spanish language. Its main characteristics are:

- The main goal is to describe the language, not to teach how to speak, or how the language reflects logical thought, etc.
- The theoretical background is taking from theoretical linguistics, more specifically, there is not a single theoretical school in the work, but it is a composition of the best experts of the Spanish language (a collective work).

As already said, the mixture of theory and description produces a very special conjuncture regarding the type of knowledge included in GDLE. Take the following definition from the GDLE:

> The noun and the modifiers that gather around the noun to specify or predicate intensional characteristics (for example, adjectives, prepositional phrases, and relative sentences) constitute the Noun Phrase. § 5.1. p 313-314

> The universal quantifiers can be sensitive to the continuous or discontinuous feature of the Noun Phrases that they select or cannot be. § 5.2.2.2. p 334

These two assertions presuppose a phrase-based description, but a rather sophisticate and elaborate schema of X bar theory. GDLE is implicitly imposing a theoretical view in its descriptions. Besides, any descriptive grammar can be described as:

- Lacking predictive power in the grammar.
- Lacking axioms or general principles.
- Lacking the definition of the syntactic structure.
- Lacking the definition of the syntax-semantics interface

In spite of this, we have chosen this grammar because of its (allegedly) neutrality and exhaustiveness, which are impossible to be found in any grammar product of linguistic enquiries framed in just one theoretical framework.

B. The knowledge elicitation strategy

Our specific objective is to elicit the knowledge implicit in this grammar to create a model of knowledge representation and inference of the Spanish language, in a framework able to integrate different types of knowledge (deep and shallow). To do that, we have followed the KADS methodology [13] of knowledge based engineering to guide the processes of knowledge acquisition and modelling. This methodology is the most adequate to manage complex and heterogeneous domains.

According to the KADS methodology, the first step is the proper analysis of the domain (in this case of the grammar itself). This analysis is known as **linguistic analysis** in the KADS terminology and consists on abstracting the basic and defining concepts of the domain (being our domain the Spanish language). This step is followed by the definition of the **expertise model**.

Linguistic analysis of the GDLE (Conceptual Model)

The first step is the proper analysis of the domain, which is constituted by the grammar itself. This analysis consists on abstracting the basic and defining concepts of the domain. Such concepts are and will be the support of the linguistic descriptions.

The result of the linguistic analysis is the conceptual model of the GDLE. The resulting conceptual model is a subjective interpretation of the data present in the GDLE, interpretation guided by the meta-descriptive statements contained the GDLE (like the description of the grammatical units as "multiform" objects, the explicit definition of the syntactic formation processes, etc.). The final analysis is the outcome of an exhaustive exploration of the GDLE.

The exploratory analysis of GDLE has as main aims the definition of the conceptual model of the domain that structures and plan the process of knowledge extraction. After the exploratory analysis, we encounter the following defining features of the conceptual model of the domain.

1. *Four types of grammatical units.* Namely: word, phrase, sentence and discourse.
2. *Multiform nature of grammar units.* Grammar units are complex objects defined according to their morphological, syntactic, semantic and pragmatic properties.
3. *Difference between Dictus* (what is said*) and Modus* (how it is said). This difference is an organizing parameter in the GDLE.
4. *Functional Units.* Grammatical Units are composed of minor grammatical units. Each grammatical constituent bears a function within the bigger grammatical units. There are syntactic, semantic and informative functions in the model.
5. *Minor Units*. There are two types of notions that although are not "physically realized", they are essential to the model, like Aspect and Time.
6. *Syntactic and Semantic Relations*. Grammar units are interrelated according to specific syntactic and semantic relations, like subordination, complementation and modification, determination, etc.
7. *Exceptions and indeterminacy in the GDLE.* Some phenomena in the GDLE are not provided with a complete description or analysis. They are presented in the form of trends, with multiple exceptions.

These seven features make up the conceptual model of the GDLE. The conceptual model of the GDLE is of paramount importance since it guides and organizes the process of knowledge extraction in a coherent way.

Expertise Model

The study of the GDLE is complemented with checking out what type of knowledge is included in the grammar. For that, we have again appealed to the dimensions that KADS considers essential to describe an expertise model, which are the following:

- Objects: defined as the concepts of the domain. These are present in the GDLE and have been recognized in the conceptual model.
- Inference rules: defined as pieces of knowledge that produces new knowledge, in our case, the rules of the grammar. The grammar rules are present in the GDLE but require a process of extraction.

- Models: defined as knowledge structures that represent complex relation in a coherent framework. Models are equivalent to the deep knowledge of a domain, the principles of linguistic theories. They are hidden in GDLE, so they need to be elicited (and subsequently we'll have to decide whether they are incorporated or not).
- Structures: defined as groups of objects and rules. They are implicit in the GDLE. They can be incomplete. They have to be explicitly expressed.
- Strategies: defined as the formalisms of representation and structure of the knowledge base. They are completely absent from the GDLE, and essential to reach a computable model of the Spanish language. There is an urgent need to define the most appropriate formalisms and representational strategies for the language model.

Objects and inference rules are visible in the GDLE, it is a matter of technique and analysis to extract and compose them. Models and structures are not visible in the GDLE; however one may think that they are implicit or hidden in the grammar. At this point, we have to choose whether the subjacent models will form part of the expertise model or not. The definition of strategies, on the contrary, is of paramount importance; since a good choice of strategies will determine the success and adequacy of the model, both from the representational (and expressivity) point of view and the computability (and operational point of view). In a way, a good choice and design of strategies will determine the final computability and plausibility of the model. Thus, the final model will be the result of (a) providing a formal structure to the knowledge organization proposed in GDLE and (b) covering the knowledge gaps regarding structures, models and strategies, which could be considered as the operative knowledge of a model.

Strategies for the knowledge extraction process

The process of knowledge extraction has been planned according to the types of grammar units (word, phrase and sentence) and not according to the different levels of linguistic analysis (structures). This has been done so because the knowledge corresponding to the grammar units is compact and coherent, while the knowledge of the linguistic levels is dispersed and scattered all over the work. Therefore, we have undergone three phases in the knowledge extraction processes: the first phase of the extraction of knowledge of different types of words; a second phase of extraction of knowledge of different types of phrases; and finally, the extraction of knowledge of different types of sentences.

Rule extraction has been done manually. Rules follow the IF THEN schema and are expressed as a conjunction of *attribute : value* pairs. For example, given the following paragraph:

> The definite article can be used as a generaliser […] The generic interpretation with the plural indefinite article is impossible.
> § 5.2.1.5. p. 327

We extract the triplets of **Table 1.**

TABLE 1. TRIPLETS.

Object	Attribute	Value
NP	interpretation	generic, specific
article	number	singular, plural
article	type	definite, indefinite

And the following rule:

```
IF NP: Espec[ Article: type: indefinite ] &
   NP: Espec[ Article: number: plural   ]
→ NP: interpretation : specific
```

Expressed in natural language as:

```
Any Noun Phrase, whose Specifier is an indefinite article in the plural number, will have a
specific interpretation.
```

The result is the definition of the objects, attributes, possible values and rules of the expertise model. We have followed traditional techniques in object-oriented analysis and design, together with the pivotal notion of roles of the grammar [14].

IV. THE KNOWLEDGE REPRESENTATION MODEL

One of the main characteristics of Artificial Intelligence is the separation of declarative and procedural knowledge. This separation constitutes one of the pillars of the maintainability of knowledge based systems. After the elicitation process, we had to distinguish between the declarative and the procedural knowledge obtained from the grammar. Thus, our strategy has been

based on the definition of a static model –constituted by the declaration of objects and relations– and a dynamic model –based on the definition of the two main processes in NLP, language understanding and language production.

Although the GDLE makes no claims about the procedural or processing aspects of language, it is suggested that both analysis and generation are based on the intervention of different types of linguistic knowledge, each one with their own rules and behaviour. The different types of linguistic knowledge are the lexical, morphological, syntactic, semantic, logical and pragmatic knowledge[1] levels of linguistic description. Each of these levels of description in the GDLE is constituted by a number of objects, rules and relations among them. Linguistic levels are independent from each other, but closely interrelated.

This situation derives in the definition of three main modules of the knowledge base of our grammatical system, which are the lexical, syntactic and semantic module. Every module is defined by the declarative knowledge and by a procedural model of behaviour.

A. Static & Dynamic Model

The **static model** is composed of the defining objects of the domain, together with general restrictions regarding their good composition, and the relations among objects that hold in the domain.

The structuring unit of the model is the Grammar Unit, which is the realization of the concept of the *multiform object* of the GDLE. Any Grammar Unit in the model (which can be an instance of Word, Phrase or Sentence) is the association of a textual representation (a string of words), a syntactic object (the structural description of the string of words) and a semantic object (the thematic and semantic representation of the propositional contents of the words string).

Besides, syntactic objects like Phrase or Sentence are composed of *syntactic functions*, like “specifier”, “subject”, “complement”, “object”, “adjunct”, etc. Such syntactic functions are conceived as roles that syntactic objects may have when they are part of a bigger syntactic object.

Figure 1 shows the relations of the static model, starting from the Grammar Unit. The relation between syntactic objects and syntactic functions is expressed by means of the *part-of* relation; whereas the relation *perform-function* expresses the relation between functions and the objects that can perform that function.

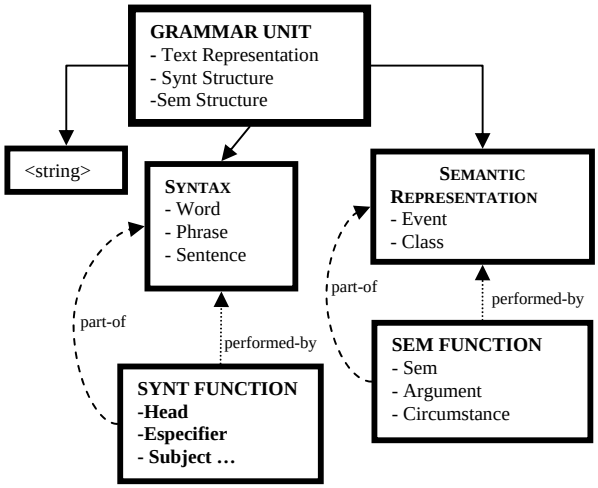

Fig 1. Static Model

An object is defined not only by its definitional properties, like attribute value pairs and components, but also by its behaviour in the domain. The objects’ behaviour of the system is defined by a number of **operations.**

Operations are defined as a grouping of rules that share objectives and functionality. The following operations are envisaged:

- Inheritance. The attribute values that define the lexical properties of heads are inherited by their phrasal and sentential projections.
- Value instantiation. This operation assigns specific values to the attributes of syntactic objects whose components and inner structure are known. These rules abound in the GDLE.
- Objects creation and transformation. These operations create syntactic and semantic objects, combine the information of two syntactic objects in one, and embed one into another. These rules are not in the GDLE but they are essential for the inference process and for the creation of complete Grammar Units.

These operations are used reversibly since they serve the analysis and generation processes. Each direction (analysis and

generation) is considered as a sequence of processes or tasks, that define with type of operations are triggered, the order in which they are triggered, how they are interpreted and the input and outputs of each subprocess. That is, the dynamic or procedural knowledge of the system is complemented with the definition of the processes for analysis and generation.

B. Missing Dimensions of the Expertise Model

Finally, the missing or implicit dimensions of the GLDE to be included in any expertise model have been tackled as follows:

1. Models (linguistic principles): We have assumed the use of argument structure, a sort of Predication Principle, constituency and use of syntactic functions, and an independent codification of the grammar from the processing direction (that is, a reversible grammar adequate for analysis and generation).
2. Structures. They are conformed by the linguistic levels described in the GDLE, that we assume. They are basic to understand and organize the static model, or the set of objects relevant of the domain.
3. Strategies. Our strategies are taken from the knowledge based literature. Our main strategies have been the following:
    a. The inclusion of two models, static and dynamic as a reflection of the division between declarative and procedural knowledge.
    b. Use of the object oriented paradigm to describe the concepts of the domain, having as main features the use of attribute: value pairs (or feature structures to describe the objects), the use of inheritance as one of the main information flows of the system, the definition of several types of operations associated to objects, and finally the use of functional objects.
    c. Definition of information flow and processes. From the execution of processes to the inheritance of values from lexical heads to phrases across the different levels of linguistic levels of description.

V. 4 A Final Note on the Architecture and Testing

It is not the object of this paper, due to space limitations, to describe on detail the architecture of the system or the development of the application, not even the testing phase, which would take several pages. However, we will include some notes about the application and the testing phase.

During the implementation of the system, it has been taken into account the peculiarities of linguistic systems, where the analysis and the generation processes are carried out by three main modules: the lexical (embedded in the syntactic representation Figure 1), the syntactic and the semantic ones. These three modules of linguistic systems are operationally independent, being only dependent at the input/output level.

One condition the representation model has to satisfy **reversibility**. That is, the rules of each knowledge base (lexicon, syntax and semantics) must serve the processes of analysis and generation with the same expression. This implies that the inference engine for each knowledge base chains the rules backwards or forwards according to the input data or initial facts. Thus, it is the data what define the process that is triggered.

The implementation of this application has been done according to a distributed architecture well known for linguistic systems developed under the framework of AI, like the blackboard architecture. This architecture have been adopted for the implementation and testing of the model, with the three main components of blackboards systems: the control module (the component that determines the direction of the data flow and the intervention of one or another knowledge source), a number of knowledge sources (lexical, syntactic and semantic ones) defined as systems composed of a set of rules and an inference engine, and a blackboard (defined as the working memory where the hypotheses selected by the control component are poured).

The developed system has been subject to a systematic process of testing with a double purpose. On the one hand, the operational validity of the model and the blackboard is tested. For example, given a syntactic tree corresponding to a sentence, check that:

- The process of generation of the semantic representation of that sentence is triggered.
- The process of generation of text is triggered.

In this way, it is checked that the blackboard works correctly (both processes are triggered); the rules of the every knowledge base are reversible; and the process of knowledge acquisition and formalization has been correctly carried out.

On the other hand, a set of test cases has been designed to prove the completion of the knowledge bases built from the GDLE. For that, it has been built a number of test cases containing a selection of grammatical phenomena enumerated in GDLE (like passive voice, subordination, complex noun phrases, etc.). For any phenomenon, it has been checked that the results were correct according to different criteria.

For the time being, execution times have not been a priority, since the goal was to prove that the reversible representation model was valid and that the methodology of knowledge acquisition from descriptive grammars was plausible and adequate.

---

[1] However, we are going to deal with just lexical, morphosyntactic and semantic knowledge at this point of development.

VI. Conclusions

We have presented another approach to the definition of grammars that veers off from the traditional "modus operandi" in NLP –based on linguistic theories- towards the use of descriptive grammars, forgotten and relegated in the formal treatment of the language, by means of the use of Knowledge Engineering methodologies. This new approach, which is supported on the use of such methodologies and therefore on the definition of a differentiated knowledge representation model and reasoning model, permits the grammar's reversibility and makes linguistic knowledge independent of the concrete application.

Besides, the use of a descriptive grammar does not require expertise or training in a given linguistic model, in contrast with grammars developed in the background of theoretical linguistics. The model presented here benefits from the larger coverage of linguistic phenomena present in descriptive grammars, as opposed to grammars derived from linguistic theories, which a more constrained account of the language coverage.

This opens a new framework in NLP for the representation of the most diverse languages; since there exists descriptive grammars for many languages (including the under-resourced ones), as opposed to the approach based on adapting the models of theoretical linguistics to a concrete application.

References

[1] M. Butt; H. Dyvik; T. Holloway King; H. Masuichi; C. Rohrer. "The Parallel Grammar Project". Workshop on Grammar Engineering and Evaluation, COLING2002, 2002

[2] J. Bresnan, (ed.). 1982. *The Mental Representation of Grammatical Relations*. Cambridge, Massachussets. MIT Press.

[3] E. M. Bender, D. Flickinger and S. Oepen. "The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars". *Procedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*. Taipei, Taiwan. pp. 8-14. 2002

[4] C. Pollard and I. Sag. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press. 1994

[5] E. Bender, D. Flickinger, J. Good and I. Sag. 2004. "Montage: Leveraging Advances in Grammar Engineering, Linguistic Ontologies, and Mark-up for the Documentation of Underdescribed Languages". Proceedings of the Workshop on First Steps for Language Documentation of Minority Languages. LREC 2004, Lisbon, Portugal.

[6] I. Bosque and V. Demonte (eds.). *Gramática Descriptiva de la Lengua Española* (3 vols.). Madrid: Espasa-Calpe. 1999

[7] R. Huddleston and G. K. Pullum. *The Cambridge Grammar of the English language*. Cambridge University press, 2002

[8] J. Cardeñosa, F. Alonso, J. Castellanos and J. García. "The application of Deep Models in Industrial Expert Systems". *Expert Systems with Applications*. Vol 2(3) pp 187-194. 1991

[9] S. Daelemans, K. De Smedt, and G. Gazdar. "Inheritance in Natural Language Processing. Special issue on inheritance: I". Volume 18 (2), Pag 205 – 218. 1992

[10] L. D. Erman, F. Hayes-Roth, V. Lesser and D. R. Reddy . "The hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty". In B. L. Webber and N. J. Nilsson, editors, *Readings in Artificial Intelligence*, pag. 349-389. Kaufmann, Los Altos, CA, 1981

[11] L. Wanner and E. Hovy. "The HealthDoc Sentence Planner". Proceedings of the 8th International Workshop on Natural Language Generation, Brighton. 1996.

[12] U. Hahn, N. Bröker and P. Neuhaus. "Let's ParseTalk: Message-passing protocols for object-oriented parsing". In H. Bunt and A. Nijholt (eds.), *Advances in Probabilistic and other Parsing Technologies*. Dordrecht, Boston, London: Kluwer. 2000

[13] G. Schreiber, B. Wielinga, and J. Breuker (ed). *KADS: A Principled Approach to Knowledge-Based System Development.* Knowledge-Based Systems Book Series, vol 11. Academic Press, London, 1993.

[14] F. Steimann. "On the representation of roles in object-oriented and conceptual modelling", Data Knowledge Engineering, Vol 35(1), pag 83-106. 2000