A Semantic Oriented Method for Conceptual Data Modeling in OntoUML Based on Linguistic Concepts[1]

Lucia Castro[1], Fernanda Baião[1], Giancarlo Guizzardi[2]

[1] NP2Tec – Research and Practice Group in Information Technology,
Federal University of the State of Rio de Janeiro (UNIRIO), Rio de Janeiro, Brazil
[2] Ontology and Conceptual Modeling Research Group (NEMO), Computer Science Department, Federal University of Espírito Santo (UFES), Espírito Santo, Brazil
{lucia.castro, fernanda.baiao}@uniriotec.br
gguizzardi@inf.ufes.br

Abstract. Conceptual data models, as means of communication, must have semantic quality. Such quality relies on the model's completeness and validity in relation to the concepts it is supposed to represent. Since the modeler acquires such concepts mostly from texts created in a natural language, a semantic-oriented linguistic approach should be adopted for building unambiguous conceptualizations. Also, the chosen modeling language must offer enough constructs for the creation of a faithful representation, like OntoUML. Such languages, however, may require a learning period that modelers hardly can afford. This paper proposes a modeling method that consists of systematic steps to promote the understanding of the concepts inherent to the domain to be modeled. The method application is illustrated in an example. Additional evaluations of the proposed method included a case study, which results indicated that it makes modeling less complex by allowing for modeling choices to be dealt with within the realm of the natural language.

Keywords: conceptual modeling, linguistics, OntoUML, semantics.

1 Introduction

Conceptual data modeling *"is by far the most critical phase of database design and further development of database technology is not likely to change this situation"* [1], and the model is a tool for intentional communication and reasoning, i.e., human activities. This paper addresses the conceptual data modeling process, which comprises two main activities: the acquisition of concepts used in the domain being modeled, and the representation of the acquired concepts in a modeling language (ML). The modeler obtains such concepts from texts produced in a natural language (NL); here, the term *text* is used in the same sense as in [2] and [3] and does not

[1] This research is funded by FAPES (45444080/09) and CNPq (481906/2009-6).

imply written material, but any product of the discourse of a community.

Discourse is described in [4] as the speech activity of an individual according to determined circumstances, and Bunge [5] defines universe of discourse as *"The collection of the possible referents of a discourse"*. In other words, the universe of discourse comprises the real-world things about which the discourse of an individual is generated through the texts he/she produces. Individual, in this case, is a community that deals with such universe in its professional activities, the members of which are said to be the domain specialists. The conceptual data model must express the domain interpretations of such specialists; the modeler must conceal personal experiences and interpretations from both the modeling process and the model itself.

The conceptual data modeling process is, then, similar to a translation activity, in terms that it consists of understanding concepts represented in a (natural) language and then representing those same concepts in a different (modeling) language. Thus, it was only natural that researchers resorted to linguistics for support in the development of methods and solutions for the modeling process, as presented in [6] and its references. However, such projects view modeling activities from the perspective of the (meta)model adopted, and linguistic concepts are used as means to support modeling decisions. Also, their work focuses on the syntactic analysis of texts, barely mentioning semantics at all; yet, translating, as well as modeling, is an activity based on "meaning", therefore handling semantics in its essence.

As with a NL to NL translation, the model is ideally expected to have the same meaning as the texts in the NL; this means to say that a conceptual model must have semantic quality. Lindlam *et al* [7] state that for a model to have semantic quality it has to be valid and complete in relation to the universe of discourse it represents. However, the modeler does not have access to such universe and his/her work has to be based on the interpretation of the domain specialists of that domain. This article presents a semantic-oriented method for conceptual data modeling that makes use of the theories of semantic types proposed by Dixon [8], as well as linguistic concepts, so as to systematically address this interpretation; this method is the result of the research presented in [9]. The ML adopted is OntoUML [10], [11], [12], [13], a well founded conceptual ML that comprises a semantically rich set of constructs. It is divided into six sections, as follows: section 2 discusses languages, both natural and modeling ones; section 3 describes the proposed conceptual data modeling method; section 4 presents a theoretical example for the application of the proposed method, section 5 discusses the method evaluation and section 6 concludes the article.

2 Languages

Bunge [5] describes language as basically a *"System of signs serving to communicate and think."* Natural language is the designation given to languages natively spoken by humans for communication. All facts and phenomena related to NLs are studied in Linguistics, which comprises semantics (study of the relations between the signs and their referents), syntax (study of the relations among signs) and pragmatics (the study of the relations between signs and the one who uses them); from these, semantics stands out since *"Understanding how we mean and how we think is a vital issue for our intuitive sense of ourselves as human beings."* [14].

The lexicon of a NL (its words) is divided into word *classes* or *parts of speech* [15] [16], that can be either *closed* (have fewer members and cannot normally be extended) or *open* (can be indefinitely extended). Open class items (nouns, adjectives, verbs and adverbs) are the ones that carry the semantic load. Dixon [5] states that the open class items of any (natural) language can be grouped into classes he names *semantic types*. All the words of a semantic type share a common meaning component and a typical set of grammatical properties, as, for instance, its association with a part of speech. The most important semantic types for conceptual modeling purposes, at least in English and other structurally similar languages, are the ones related to concrete-referenced nouns, since this is the class of words that name types of things. Dixon [5] groups such nouns as follows: Animate (in this case, animals), Human and its subclasses (Kin, Rank and Social Groups), Parts (body and others) and Inanimate and its subclasses (Artefacts, Celestial and Weather, Environment and Flora). Verbs are important for establishing relations between concepts. Semantic types associated with verbs are classified as Primary (*"refer to some activity or state; verbs that can make up sentences by themselves"*) and Secondary (*"those providing semantic modification of some other verb"*). Semantic Types associated with Adjectives, on the other hand, are divided in 11 subclasses: Dimension, Physical Property, Speed, Age, Colour, Value, Difficulty, Volition, Qualification, Human Propensity and Similarity.

Apart from using their NLs for communication, men have been creating abstractions (i.e., building models) of real-world things in a way to understand and cope with reality [17]. For models to be understandable and useful to a community, they must be created from a system of symbols and connecting rules (grammar) known to all members of that community. Such systems are MLs, which are artificial languages also used for communication and to help reasoning, through the creation of models instead of texts. This work adopts OntoUML as ML, that, due to its underlying foundational ontology (UFO) [10] [13], provides constructs enough to allow for the creation of semantically accurate models. However, using such a language can present a problem since it requires a deeper knowledge of the philosophical concepts that are the bases for its constructs meanings, and a training period that most modelers cannot afford.

Different from NLs, ML**s** do not provide a lexicon; consequently, the translation between a NL and a ML must be done through the comparison between NL constructs (here, semantic types) and the constructs of the ML (the NL sign representing the concept being modeled appears in the model as the label of a construct). Bunge [5] defines *construct* as *"a concept, proposition, or set of propositions, such as a classification, a theory, or a moral or legal code"*; both natural and modeling languages have meaningful constructs. For instance, each of Dixon's semantic types [8] can be considered a construct, as well as each category described in the ML. Constructs are defined in terms of meta-properties, which must be compared during the modeling process so that the meaning restrictions imposed by the NL constructs are present in the model, reflected in the ML construct used in each representation.

3 The method

This paper proposes a method for the creation of conceptual data models in

OntoUML. The main goals are: to provide means for modelers to understand the concepts presented in the texts produced by domain specialists; to prevent modelers' from representing their own interpretation of the domain, instead of the specialists'; to allow for modeling decisions to be made within the realm of the NL, so that even modelers with little experience in OntoUML are able to create accurate models; to help creating models that have semantic quality, by ascertaining that the representations are valid and complete; and to provide means for this semantic quality to be maintained through time. The proposed method consists of six steps:

Step 1 – Breaking the text into kernel sentences - The modeler decomposes the NL texts produced by domain specialists into kernel sentences. Kernel sentences are affirmative, active sentences that do not have co-ordinate or subordinate clauses [18] [19]. They form the deep structure (meaning) of a text, whereas the surface structure (form) of the text is the result of transformations applied to the deep structure (e.g., identical subject suppression and passivization [19]). To extract the kernel sentences from a text, one should reverse such transformations. A simple example could be the sentence *John went to the beach and was taken home afterwards.* The sentence includes two clauses co-ordinated by the conjunction *and. John* is suppressed in the second clause, since it is the subject in both. Also, we know that someone took *John* home after he left the beach. The technique, thus, for "breaking" complex sentences is looking for co-ordinate and subordinate conjunctions and understanding how they relate clauses, identifying suppressed subjects, and converting sentences from passive to active voice, whenever applicable. When a resulting active voice sentence does not have an explicit subject, the word "someone" should be used as substitute (this specifies points to be clarified with the user in Step 2). So, for the example above, we could have two kernel sentences: *John went to the beach* and *Someone took John home afterwards.* Kernel sentences must be arranged in a numbered list in the order they appear in the text, so that reading the list is like reading the text itself.

As the modeler decomposes the text into simple sentences, he/she may find that pieces of information are missing or find ambiguities that will have to be explained by the domain specialists. One way of spotting missing information is to identify the verb semantic types and their related semantic roles and make up questions according to those roles. For instance, the verb *give* imply that something (*gift*) that belonged to someone (*donor*) will now belong to someone else (*recipient*). Table 1 presents some of the semantic types for verbs [5], their related semantic roles and the questions that might be asked in order to discover missing information. The modeler should write a list with all questions and doubts, in the same order as they appear in the text; this list, as well as the list of simple sentences, is the output for Step 1.

Table 1. Questions for spotting semantic roles

Semantic Type	**Semantic Role**	**Questions**
Affect	Agent	***Who*** *<verb> <Target> with <Manip>?*
	Target	*<Agent> <verb>* ***whom/what*** *<Manip>?*
	Manip	*<Agent> <verb> <Target>* ***with what****?*
Giving	Donor	***Who*** *<verb> <Gift>****to*** *<Recipient>?*
	Gift	*<Donor> <verb>* ***what to*** *<Recipient>?*
	Recipient	*<Donor> <verb> <Gift>* ***to whom/what****?*

Semantic Type	**Semantic Role**	**Questions**
Corporeal	Human	***Who*** *<verb> <Substance>?*
	Substance	*<Human> <verb>* ***what****?*
Competition	Competitor	***Who*** *<verb>?*
	Activity	***Competitor*** *<verb> <Activity>?*
Social Contract	None	***Who*** *<verb>* ***who****?*
Using	None	***Who*** *<verb>* ***what****?*

Step 2 – Clearing doubts - The modeler must then meet with domain specialist(s) and clear all doubts. According to the answers provided by the domain specialists, the modeler updates the list of simple sentences, explicating previously unknown subjects, and eliminating synonyms and ambiguities.

Step 3 – Identifying signs - The modeler must identify the conceptually significant NL signs present in the sentence list. In English, as well as in other similarly structured languages, such symbols will be nouns, verbs and adjectives. Such signs must be organized in a table with columns for the subject, the verb and the objects of each simple sentence – each row of the table will represent a simple sentence.

Step 4 – Linking signs to Semantic Types - The modeler must associate each of the identified signs with one of the semantic types. As semantic types are not mutually exclusive, the modeler must be careful so as to make the association that is applicable in that specific context or domain. The output for this phase is the table of signs; each row presents the sign and the semantic type to which it was associated.

Step 5 – Mapping Semantic Types to OntoUML constructs – In this step, the modeler systematically identifies a preliminary set of OntoUML constructs that will be needed to model the concept each sign previously identified represents. This is conducted by applying the mappings defined in [9]; some mapping examples are illustrated in Table 2. This mapping tends to be fairly stable and, as such, it can be organized in a table that can be accurately used in most situations.

Table 2. Semantic Types to OntoUML constructs Mapping

Semantic Type	**OntoUML Construct**	**Semantic Type**	**OntoUML Construct**
Animate	*Kind*	Social Group	*Kind*
Human	*Kind*	Part	When the part is a component, *kind* When the part is an ingredient, *quantity* When the part is a member, *kind* When the part is a sub-collection, *collective*
Kin	*Role*	Inanimate	*Kind*
Rank	*Role*	Artefact	*Kind*

Step 6 – Creating the model - Once the semantic types have been mapped to the ML constructs, the model can be created and taken to the domain specialist for validation, before the final model is produced.

4 Example

In their seminal conceptual modeling book [1], Battini *et al* provide exercise case studies for students. We have selected a small excerpt of the text for one of such exercises (pp 268--269) for our example modeling.

“In the library of a computer science department, books can be purchased both by researchers and by students. Researchers must indicate the grant used to pay for the book; each student has a limited budget, which is fixed each year by the dean of the college.”

The list of simple sentences produced in step 1 is:

0. In the library of a computer science department
1. Researchers can purchase books
2. Students can purchase books
3. Researchers pay for books with grants
4. Researchers must indicate the grant used to pay for the book
5. Each student has a limited budget
6. Each student pays for books from their budget
7. The Dean of the college fixes students’ budgets every year

The question produced in Step 1 is answered in Step 2 as follows:

Q: When you say “grant”, do you refer to the amount of money or to a document, like a grant report, or a grant certificate?

A: It refers to the amount of money (library budget).

Table 3 presents the signs identified in each sentence.

Table 3. List of signs

Sentence	Signs			
0	Library	Comp. Sci. Dept.		
1	Researcher	Purchase	Book	
2	Student	Purchase	Book	
3	Researcher	Pay	Book	Grant
4	Researcher	Indicate	Grant	Book
5	Student	Have	Limited Budget	
6	Student	Pay	Book	Budget
7	Dean	Fix	Student	[yearly] Budget

The next step is the association of identified signs with semantic types. Table 4 presents the list of signs and the rationale behind their associations.

Table 4. List of signs and their associations with semantic types

Sign	Semantic Type
Computer Science Dpt.	Noun phrase that refers to a division of an institution, i.e., it has a concrete reference, is related to humans and is a **Social Group**.
Library	Noun that also refers to a division of an institution; also has a concrete reference, is related to humans and is a **Social Group**.
Researcher	Sign refers to a human but qualifying the person according to a position and/or responsibility; the semantic type should be **Rank**.
Purchase	Purchase is a Primary A verb, of the type **Giving**, i.e., one that always involves 3 semantic roles: a donor, a donated thing and a recipient.
Book	An object (concrete and inanimate) produced by men, thus, an **Artefact**.
Student	Sign refers to a human but qualifying the person according to a position and/or responsibility; the semantic type should be **Rank**.
Grant	Sign refers to an amount of money given by an organization for a particular purpose; a nominalization of the verb *to grant*, it’s meaning relates a Primary A verb of the type **Giving**.
Budget	Sign refers to an amount of money set aside for a particular purpose; a nominalization of the verb *to budget*, it’s meaning relates to a Primary A verb of the type **Giving**, since the Dean fixes the amount of money a Student has at his/her discretion, and this procedure is repeated every year.
Dean	Sign refers to a human but qualifying the person according to a position and/or responsibility; the semantic type should be **Rank**.

The modeler then must map semantic types to OntoUML constructs, following table 2. For example, a Social Group is mapped to Kind, Rank to Role, Giving to Relator and Artefact to Kind (detailed rationale beyond this mapping is explained in [9] and [20]). Finally, the modeler creates an OntoUML model and validates it with domain specialists. Figure 1 presents the produced version of our example model.

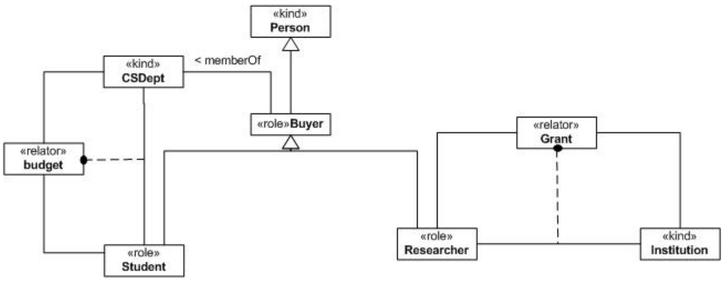

Fig. 1. OntoUML example model

5 Method Evaluation

The method presented in this paper was evaluated through a case study and an experiment. Due to space restrictions, details about this evaluation are presented in [20]. The results of the experiment showed that modelers found it easier to discuss concepts in terms of NL constructs, and that the model produced according to the proposed method was complete and valid, i.e., the model had semantic quality.

6 Conclusion

This work proposed a method for conceptual data modeling in OntoUML that is based on linguistic analysis and on the semantic types theory proposed by Dixon [5]. We developed the mapping of each of those semantic types to the constructs of a well-founded ontological ML, OntoUML. MLs differ from NLs in that the meaning of representations do not come from signs but from constructs; thus, the modeler must compare NL constructs (semantic types) to the ML ones, from the meta-properties inherent to each of them. One quality trait of the produced model relies on its semantic equivalence to the descriptions provided in the NL. The use of ontological languages to achieve semantic quality is not a novelty and the semantic gain of an OntoUML model over a correspondent ER one is evidenced in [12]. However, the semantics of such language constructs is much less intuitive for the modeler than the semantics of the constructs of his NL; thus, discussing concepts and understanding the metaproperties that apply to them is much easier if done in the NL. The method proposed in [20] uses linguistics to achieving semantic quality in conceptual models, not only by the application of semantic principles but also by providing a systematized list of activities to achieve this goal.

The outputs of each step of the method form a record of the modeler's rationale throughout the modeling process; this is important for keeping the semantic quality of the created method. NLs are essentially ambiguous, and provide several ways of saying the same thing; also, NLs are in constant evolution and semantics are affected by it, i.e., the meaning of signs may change with time. MLs, on the contrary, need to provide for unambiguous representations of concepts; and models are static representations that may provide erroneous information as time passes. Consequently, recording the reasons why constructs and signs were chosen to represent a concept is a way of maintaining the semantic quality: people who read the model in the future can use the documentation created during the modeling process to understand such choices and the semantics behind them.

References

1. Batini, C., Ceri, S., Navathe, S.: Conceptual Database Design. Benjamin/Cummings (1992)
2. Eco, U.: Semiotics and the Philosophy of Language. Indiana Univ. Press (1984)
3. Koch, I.: Introdução à Linguística Textual. WMF Martins Fontes (2009) (In Portuguese)
4. Bechara, E.: Moderna Gramática Portuguesa. Nova Fronteira (2009) (In Portuguese)
5. Bunge, M.: Philosophical Dictionary. Prometheus Books, Amherst (2003)
6. Castro, L., Baiao, F., Guizzardi, G.: A Survey on Conceptual Modeling from a Linguistic Point of View. Technical Report, Rela Te-DIA (2009)
7. Lindlam, O., Sindre, G., Sølvberg, A.: "Understandig Quality in Conceptual Modeling", IEEE Software, v. 11, n. 2 (Mar), pp. 42-49 (1994)
8. Dixon, R. M. W.: A Semantic Approach to English Grammar. Oxford University Press, Oxford (2005)
9. Castro, L., Baião, F., Guizzardi, G.: A Linguistic Approach to Conceptual Modeling with Semantic Types and OntoUML. In: Intl Workshop on Vocabularies, Ontologies and Rules for the Enterprise (VORTE 2010), Vitoria. EDOC 2010 Workshops (2010).
10. Guizzardi, G.: Ontological Foundations for Structural Conceptual Models. CTIT (2005)
11. Benevides, A. B., Guizzardi, G.: A Model-Based Tool for Conceptual Modeling and Domain Ontology Engineering in OntoUML. In: Filipe, J., Cordeiro, J. (eds) ICEIS 2009. LNBIP, vol. 24, pp 528--538, Springer, Heidelberg (2009)
12. Guizzardi, G., Lopes, M., Baião, F., Falbo, R.: On the Importance of Truly Ontological Distinctions for Ontology Representation Languages: An Industrial Case Study in the Domain of Oil and Gas. In: Holpin, T., Krogstie, J., Schmidt, R., Soffer, P., Ukor, R. (eds) BPMDS 2009 and EMMSAD 2009. LNBIP 29, pp 224--236, Springer, Heidelberg (2009)
13. Benevides, A. B., Guizzardi, G., Braga, B. F. B., Almeida, J. P. A.: Assessing Modal Aspects of OntoUML Conceptual Models in Alloy. In: Heuser, C., Pernul, G. (eds) ETheCoM 2009. LNCS, vol. 5833, pp 55--64. Springer, Heidelberg (2009)
14. Jackendoff, R.: Foundations of Language. Oxford University Press, Oxford (2002)
15. Greenbaum, S.: The Oxford English Grammar. Oxford University Press, Oxford (1996)
16. Quirk, R., Greenbaum, S.: A University Grammar of English. Longman, London (1973)
17. Schichl, H.: Models and History of Modeling. In: Kallrath, J.: Modeling Language in Mathematical Optimization. Pp 25--36, Kluwer Academic Publishers, Norwell (2004)
18. Chomsky, N.: Aspects of the Theory of Syntax. MIT Press, Cambridge (1965)
19. Chomsky, N.: Syntactic Structures. Mouton de Gruyter, New York (2002)
20. Castro, L.: Abordagem Linguística para a Modelagem Conceitual de Dados com Foco Semântico, MSc Dissertation, Unirio, Rio de Janeiro (2010) In Portuguese