A Linguistic Approach to Conceptual Modeling with Semantic Types and OntoUML

Lucia Castro, Fernanda Baião
NP2Tec-Research and Practice Group in Information Technology
UNIRIO, Rio de Janeiro, Brazil
{lucia.castro, fernanda.baiao}@uniriotec.br

Giancarlo Guizzardi
NEMO-Ontology and Conceptual Modeling Research Group
UFES, Vitória, Brazil
gguizzardi@inf.ufes.br

Abstract— **The process of conceptual modeling involves the acquisition of concepts (and of the signs that represent them) used in the Universe of Discourse (UoD) being modeled, and the creation of the model (as a concrete artifact) according to a modeling language grammar. The knowledge about the UoD is obtained from a variety of sources, all of which are mostly expressed in a natural language. It is correct to say that conceptual modeling is much similar to language translation i.e., identifying concepts that are represented by signs of a language, and then representing those same concepts in a different language. Also, the semantic quality of the resulting model (translation) is directly affected by the modeler's (translator's) understanding of the source material. As so, conceptual modeling activities can benefit from an analysis carried out from a linguistic point of view, as well as from the use of a modeling language which constructs allow for a representation that is semantically equivalent to the natural language original descriptions. This work proposes a linguistic approach to conceptual modeling based on the notion of semantic types, and on the use of OntoUML as a modeling language. The proposed approach is illustrated in an example.**

Keywords: conceptual modeling, OntoUML, linguistic approach

I. Introduction

Conceptual 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]. In fact, the results of an empirical study on enterprise conceptual modeling [2] reveal that *"...the vast majority of modeling teams are sketching and not using CASE or CAD tools."* This is a consequence to the fact that a conceptual model is also a tool for intentional communication and reasoning (human-centered activities), as opposed to an artifact strictly constructed for technical activities such as system development, database interoperability automation, inference mechanisms on ontologies, as many tend to think.

The process of conceptual modeling involves two main tasks: the first one comprises the acquisition of concepts related to the Universe of Discourse (UoD) being modeled (aka conceptualization [7]), along with the identification of the natural language signs used to represent such concepts; and the second involves the representation of the acquired concepts according to the grammar of a modeling language (ML), i.e., the creation of the actual model. The knowledge about the UoD is obtained in a variety of ways, like from interviews with users, reports and functional documents pertaining to that environment, from observation of the group routine, etc. However, no matter the source of information, the knowledge about the scenario to be modeled is, in most cases, passed to the modeler in a natural language (NL).

To develop a conceptual model, the modeler must identify conceptual elements present in the UoD descriptions, understand how they relate to each other and then represent both conceptual elements and their inter-relationships in a ML [3]. It is correct to say that the conceptual modeling process is a translation activity, i.e., identifying concepts that are represented by signs that belong to a language, and then representing those same concepts with signs that belong to a different language.

Defining the process of conceptual modeling as a translation activity is not a novelty and, consequently, it was only natural that researchers resorted to linguistics for support in the development of methods and solutions for the modeling process, as illustrated by the research works presented in [4] and its references. However, such projects view the modeling activities from the perspective of the (meta)model, and linguistic concepts are used, at best, as a means to support modeling decisions. For example, in [5] Chen proposes 11 rules for the translation of English descriptions into ER models; rule 1 states that *"A common noun in English corresponds to an entity type in an ER diagram."* - it is true that an entity type is named after a concept that is represented by a common noun but, considering that all nouns that are not proper are common, should a common noun always correspond to an entity type in an ER model? Also, they are restricted to syntactic constructs (parts of speech) that are very broad in terms of meaning. This research proposes that the linguistic constructs should convey semantic properties and not just syntactic ones.

As with a NL to NL translation, the resulting conceptual model is ideally expected to provide the same reading as the texts or representations built in a NL, which can be described as semantic accuracy. To accomplish this task, the modeler also needs to choose a ML that is as expressive as its natural counterpart, and that provides the means for the creation of clear and sound representations of the modeled domain.

In order to explicitly and systematically take semantics into account, as well as adequately represent higher quality conceptual models, this work proposes the use of Semantic Types [6], and shows their relation to the constructs of the OntoUML ML, which is described and documented in [7], [8], [9] and [10]. We also sketch a sequence of steps that should be carried out by a conceptual modeler who follows our approach. It is divided in six sections, as follows: section II discusses languages, both natural and modeling ones; section III describes the translation or mapping between natural and ML; section IV describes the proposed approach; section V proposes a modeling example, and section VI concludes the article.

II. LANGUAGES

Bunge [11] states that a language is a *"System of signs serving to communicate and think."*, as do most dictionaries. Any language is a system intentionally used by humans for communication and reasoning, and composed of signs which meanings are determined by opposition to one another. This definition of language applies both to natural and modeling ones; however, there are characteristics that are particular to one or the other that must be understood.

A. *Natural Languages*

Natural language is the designation given to languages natively spoken by humans in order to transmit information, express emotions and requests, for social interaction, and poetic and creative expressions. McWhorter [12], in his presentation of the history of languages, describes how one single language - that was first used approximately 150,000 years ago - evolved to form what are now the more than 6,000 NLs known in the world. According to the reality faced with by each of its groups of users, who migrated from East Africa to the rest of the world, the original language evolved until the separate *dialects* morphed into completely diverse languages. That means to say that, although a NL presently learned by human beings may seem a static structure, it is a dynamic entity in slow but constant change and evolution, in order to reflect and meet its speakers' communication and reasoning needs.

All facts and phenomena related to NLs are studied in Linguistics, which is, according to Saussure [13], a branch of Semiotics, the study of signs and symbols in general. Linguistics focuses on the study of the linguistic sign, and 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 meaning is the "holy grail" not only of linguistics, but also of philosophy, psychology and neuroscience, to say the least, since *"Understanding how we mean and how we think is a vital issue for our intuitive sense of ourselves as human beings."* [14].

Saussure, also in [13], describes the linguistic sign as the union of a *signifier* (an acoustic image or a articulated word) and a *signified* (a concept). Ullmann [15] later presents not two but three elements of meaning: a *symbol* or *sign* represents a *concept* and refers to a *thing* in the real world, or the referent; the concept is an abstraction of the thing. The elements of meaning are commonly presented as the vertices of a triangle, known as "semiotic triangle" or "Ullmann's triangle" or even "Ogden and Richard's triangle of meaning", which can be seen in figure 1.

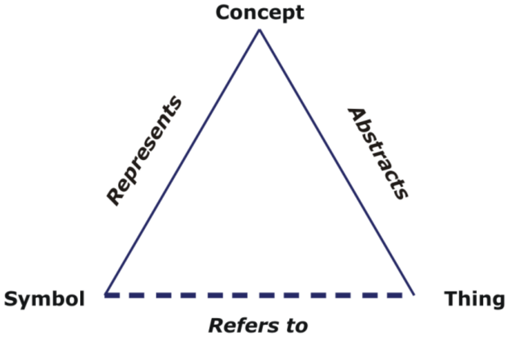

Figure 1. The semiotic triangle

The lexicon of a NL, or its vocabulary, is divided into *word classes* or *parts of speech*, which are sets of words grouped according to notional, morphological and grammatical criteria [16]. These classes themselves are further grouped into two other classes: the *closed-system items*, or *closed classes*, and the *open-class items*, or *open classes*. Closed classes are the ones that have fewer members and cannot normally be extended by the creation of new items; they include determiners, prepositions, pronouns, conjunctions and numerals. Open classes, on the other hand, are the most numerous, and the ones that can be indefinitely extended; these are nouns, adjectives, verbs and adverbs [16] [17]. Open class items are the ones that carry the semantic load, and therefore are the focus for conceptual modeling.

The first step in conceptual modeling is the acquisition of the concepts and symbols related to a certain UoD from documents and/or statements written or uttered in a NL. As so, in this first step the modeler must focus on understanding the elements present in the triangle above, i.e., the identification of the signs used in that context and the understanding of their relations to the concepts and the reality they abstract. In other words, this is a semantic-oriented activity. Only then can a modeler try and represent the same Universe in a ML.

B. *Modeling Languages*

Men have been creating abstractions of real-world things since Stone Age, as a way to understand and cope with reality [18]; thus, it is right to say that men have been building models ever since. According to Guizzardi [7], *"A model is an abstraction of reality according to a certain conceptualization."*; however, for models to be understandable and useful to a community, they must be created according to a known system of symbols and connecting rules. Such systems are MLs, which are artificial languages that, as their natural counterparts, are also used for communication and to help reasoning but, in this case, through the creation of models. In terms of information systems, conceptual MLs started being formally defined in the 1970's ((E)ER, CSL, UML, DAML, OWL).

While NLs allow for several diverse ways to express the same content, and even for the creation of messages that, although grammatically correct, are not precise in meaning (and this is a very desirable trait since it favors creativity, as expressed in poetic and literary usage, besides jokes and puns), a ML must be strict and prevent ambiguity. In fact, Guizzardi [7] states that such languages must be *lucid* (offers constructs enough to represent diverse concepts), *sound* (does not provide constructs which interpretation is not clear for the modeler, leading to wrong or diverse usage), *laconic* (does not allow for an entity to be represented in more than one way) and *complete* (provides for the representation of all possible concepts). As so, not only a systematic approach must be adopted in conceptual modeling, but also the choice for a language which constructs and grammar prevent ambiguity and non-semantic constructions are key in the development of good and efficient conceptual models. In this work, we adopt a language with such characteristics, i.e., an ontologically well-founded and semantically rigorous conceptual ML named OntoUML.

III. TRANSLATION

If there are 6,000 NLs, it is correct to say that there are at least 6,000 ways of expressing a certain message. It is also safe to say that native speakers of different NLs may need to communicate and that, for such communication to occur, it may be necessary that the message to be exchanged between them be translated from one NL to another.

A. *Translation between natural languages*

Mounin [20] discusses problems related to the translation activities between NLs. He introduces the concept of *semantic field*, or *area of meaning*, to which a group of words would belong, and within which the meaning of each word is determined in opposition to the meaning of the others. The concept of habitation, for instance, can be generically represented by the sign *house*, but can also be specialized into the concepts represented by *apartment*, *bungalow*, *shack*, *mansion*, *manor*, or even *palace*, according to differentiation and comparisons among themselves; all these signs belong to the area of meaning that comprises the representations for dwelling places. The concept of semantic field is important for translation activities because it provides a way to deal with the fact that each linguistic system (or language) presents a way of perceiving the world that differs from the other ones; that is to say, when translating a message, the sign of the target language must be chosen from among the ones that belong to the same semantic field as the sign in the source language. Figure 2 represents the translation process between NLs.

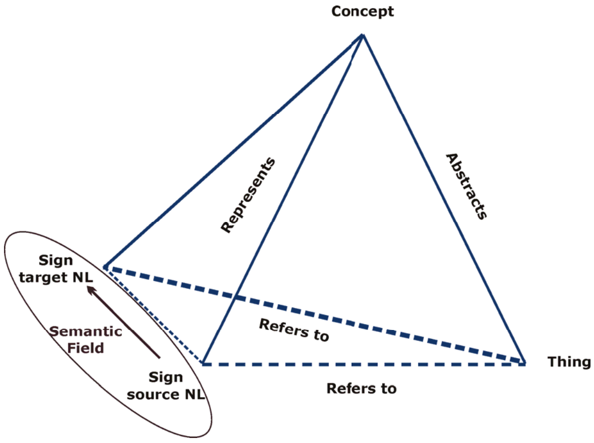

Figure 2. The translation process

B. *Translation from natural language to modeling language*

Although the process of translating NL concepts into a conceptual model implies comparison techniques, as with the translations between two NLs, there are two main differences that must be taken into account. First, the modeler and the domain experts need to communicate in the same NL. Second, the abstraction level of a ML is such that only more abstract NL concepts can be expressed equivalently in the ML; i.e., comparison is not done between equivalent semantic fields but between NL constructs and the abstract constructs of the ML. (The NL sign representing the concept being modeled appears in the conceptual model as the label of a construct.)

To illustrate this comparison with an unbiased example, we will use the word *wasi* from Quechua, a native American language spoken in the Andes. Saying that it is a concrete-referenced sign and that it refers to an artifact, narrows down the meaning possibilities for it, that definitely cannot be flower, or red, or hunger, for example. Having a concrete reference, being inanimate, and referring to an artifact are properties that define the semantic class to which this word belongs, and narrows down its meaning possibilities. The ML must have constructs that convey the same defining properties in order to provide for the semantic equivalence with the NL description. The NL constructs this research is based on are Dixon's Semantic Types [6].

C. *Semantic Types*

In [6] Dixon presents a novel, meaning oriented grammar of English, which principles can be applied to other NLs. He 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. In fact, semantic types are semantic specializations of parts of speech, in this case, nouns, verbs and adjectives. A full discussion of all Semantic Types proposed by Dixon is outside the scope of this work but here are some explanations that are necessary for the understanding of the proposal.

The most important semantic type for conceptual modeling purposes, at least in English and other structurally similar languages, are the ones related to nouns, since this is the class of words that name beings and things. For instance, the ones that have a concrete reference can be 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). Figure 3 presents the semantic types linked to nouns.

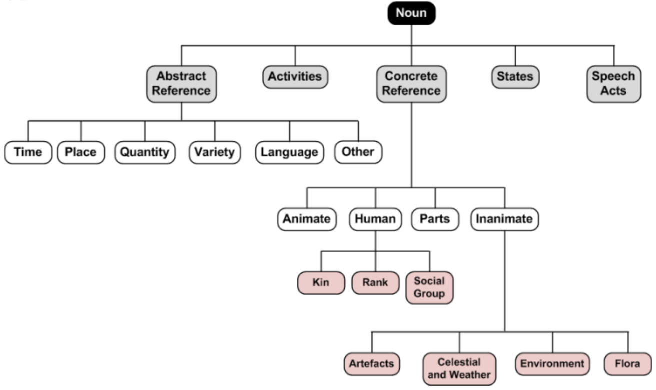

Figure 3. Semantic types associated with Nouns

Concrete reference nouns are mostly the focus on conceptual data modeling, since they represent concepts that abstract real-world things about which data is supposed to be stored. They are the ones which referents are beings or things that exist independently in the world [22]; or the ones that *"refer to entities that are typically perceptible and tangible"* [16]. Either way, the independent existence, perceptibility and tangibility imply the possibility of individuation and identification of such referents. The *Concrete reference* semantic type specializes into four other semantic types:

- *Animate* – semantic type that groups concrete reference nouns which referents are living beings that are members of the kingdom Animalia but are not human. Examples include all non-human animals, like *dog, cat, horse,* etc.
- *Human* – semantic type that groups concrete reference nouns which referents are living beings that are members of the kingdom Animalia and are human. It further specializes in:
    - *Kin* – semantic type that groups humans, and denote a position distinction of one element in relation to others in a family group. Since a same element may have different positions in different family groups, either biological or legal, (a same person can be a father, a son, a nephew, a husband, a son-in-law etc), *Kin* references cannot be mutually exclusive; however, such positions are not interchangeable within the same biological family group, i.e., a son cannot exchange his position with his father, for instance.
    - *Rank* – semantic type that groups humans, and denote a position distinction of one element in relation to others in a social group. Since a same element may have different positions in diverse social groups (a man can, at a certain point of his life, be a goalie, a teacher, a manager, and a volunteer coach), *Rank* references cannot be mutually exclusive and they can be interchangeable within the same social group, e.g., a *goalie* can become a *stopper* in the same team if needed be.
    - *Social Group* – semantic type that groups sets of humans, which are seen and identified as a whole. A company is an example of a *Social group*.
- *Parts* – semantic type that groups concrete reference nouns, which referents are parts, i.e., that are seen as the *part* in a *part-whole* relation with other beings; such parts can be corporeal or not.
- *Inanimate* – semantic type that groups concrete reference nouns, which referents are inanimate things. It further specializes in:
    - *Artefacts* – semantic type that groups inanimates that are created or made by men, i.e., things that are not present in nature, like *book* , for instance.
    - *Flora* – semantic type that groups inanimates which referents are living beings that are members of the kingdom Plantae, like *tree* or *daisy.*
    - *Celestial and weather* – semantic type that groups inanimates which referents are celestial bodies, like *sun* and *moon*, or related to weather or climate, like *rain* or *wind*.
    - *Environment* – semantic type that groups inanimates which referents are related to the environment and mostly members of the mineral kingdom, like *water* and *gold*, for instance.

Most referents of *concrete reference* nouns are countable but some of them that are non-countable; in this case, their referents are seen and/or perceived as a mass. They are generally minerals and belong to the *environment* semantic type, like *water* and *gold*, as mentioned above.

In terms of conceptual modeling, verbs are important when they establish relations and links between concepts, that can be either *agents* (the ones who act, or syntactic *subjects*) or *patients* (the ones that receive the impact of the action, or syntactic *objects*). Dixon calls those *Primary* verbs, i.e., they refer to an activity or state, and can form sentences by themselves, provided that agents and patients are properly stated. Primary verb type subdivide into A (verbs that must have nouns or noun phrases as both agent and patient) and B (can take clauses as agents and/or patients). For example, in the sentence *John threw snowballs at Helen.*, the verb (*throw*) is primary A, since its agent (here, *John*) and patient (here, *snowball* and *Helen*) positions cannot be occupied by another action (or clause). In the sentence *Lucy enjoys talking to John.*, however, the verb (*enjoy*) is a primary B one since it can take an action (*talking to John*) as patient (or object). Primary A verbs subdivide in *Motion, Rest, Affect, Giving, Corporeal, Weather, Competition, Social Contract, Using* and *Obeying*; Primary B verbs further classified as *Attention, Thinking, Deciding, Speaking, Liking, Acting* and *Happening.* Such specializations are based mostly on the number and types of semantic roles involved in the actions. For example, *giving* verbs involve a *Donor* (agent), a *Gift* and a *Recipient* (patients). All other verbs are *Secondary* verbs, i.e., the ones that appear together with other verbs, semantically modifying them. An example is *begin*, as in the sentence *Mary began cooking dinner at six.*

Finally, Semantic Types associated with Adjectives are divided in 11 classes: Dimension, Physical Property, Speed, Age, Colour, Value, Difficulty, Volition, Qualification, Human Propensity and Similarity.

The semantic types proposed by Dixon [6] are present in all NLs, but the word classes to which they are associated may vary. Also, a word may be linked to different semantic types in the same NL, depending on the context in which it is used: the word *book*, for instance, can be a noun, when referring to a number of written sheets of paper bound together (an *artefact)*, or a verb, when it refers to the act of reserving something (a *giving* type verb). The present work takes semantic types as the NL constructs to which the ML ones must be compared when a model is built.

D. OntoUML

OntoUML is a ML based on a revision of a portion of UML, and which constructs derive from the Unified Foundation Ontology (UFO) [7] [10]. A foundational ontology is taken here as *"a domain-independent common-sense theory constructed by aggregating suitable contributions from areas such as descriptive metaphysics, philosophical logics, cognitive science and linguistics."* [19]. Based on philosophical concepts, and starting from *universals*, the UFO describes real world categories and, as so, provides OntoUML with the basis for the creation of semantically accurate models.

Although a complete explanation of UFO concepts and the OntoUML constructs is beyond the scope of the present work, some of them will be briefly discussed here, as a means for the understanding of the proposal presented later. The first step would be to define that *universals* are properties that determine classes of all the things; *sortals* are *universals* that determine classes of things and beings that have an identity and individuation principle. *Sortals* can be *substance sortals* (properties that provide identity principles to its instances and that necessarily apply to such instances) or *phased-sortals* (properties that apply to its instances contingently).

- *Kind* – properties (*substance sortal*) that rigidly (all members in every world must have) determine classes of complex beings or things that are relationally independent, and that can be clearly identified; the beings or things that are modeled by this construct can be animals (Dog, Cat, Person), plants (Tree), artifacts (Chair, Book, Television), and institutional agents (Organization, Football Team)
- *Quantity* – properties (*substance sortal*) that determine classes of mass substances, like Gold, Water, Clay.
- *Collective* - properties (*substance sortal*) that determine classes of collectives, i.e., collections of members of a class of beings or things (as the ones determined by the *Kind* construct) seen and perceived as a uniform structure (e.g., a Forest, a Flock, a Pack [of lions] a Pile [of bricks]).
- *Phase* – properties (*phased-sortal*) that determine classes of partitions of beings or things determined as *Kind*. The classes determined by *Phase* are stages in the existence of a being; they are mutually exclusive and depend on the intrinsic properties of an individual. A good example would be Caterpillar and Butterfly that are partitions (*Phases*) of a Lepdopterum (*Kind*).
- *Role* – properties (*phased-sortal*) that determine classes of relationally dependent roles of beings or things determined as *Kind*. The classes determined by *Role* are not mutually exclusive and depend on the extrinsic properties of an individual. A good example would be Student that is a role of a Person (*Kind*). [7]

The definitions provided here have been simplified as to fit the purposes of this work. The UFO, as well as OntoUML, involves many other philosophical concepts that must be understood for the creation of correct and precise models. OntoUML constructs can be Concepts (*Category, Collective, Kind, Mixin, Mode, Phase, Quantity, Relator, Role, RoleMixin, SimpleDataType, StructuralDataType,* and *Subkind*) or Relations (*Characterization, ComponentOf, Datatype Association, Derivation, Formal, Generalization, Material, Mediation, MemberOf, SubCollectionOf,* and *SubQuantityOf*).

E. Semantic Types and OntoUML constructs

As mentioned before, the translation from a NL to a ML involves the comparison of the constructs of the first to the ones of the latter in order to build a semantically accurate model. This comparison is based on the defining (meta)properties of one and the other. For instance, in our theoretical example, one of the signs is *Book.* Book is an *Artefact*, i.e., it is a noun which referent is concrete, countable, and inanimate; to represent a Book, we must find an OntoUML construct which properties define a class of things that exist independently, that are countable, are inanimate and a made by man – an *Artefact* should be mapped to a *Kind.* Such comparison is not trivial, however, and there is not a one-to-one clear unique possibility – languages are culturally dependent and subjective; also, the classification of signs according to semantic types depends on the Universe being modeled.

For the case of *concrete reference* nouns, we have that their referents are all determined by *sortals.* According to the definition and examples provided above, most of the *concrete reference* semantic types are mapped to the *Kind* (classes of existentially independent, determinately identified and individuated, rigidly determined beings or things) OntoUML construct; this shows that the semantic types proposes a more detailed differentiation of such classes. The same happens with the semantic types *Kin* and *Rank* that are mapped to the *Role* OntoUML construct; however, the *Phase* construct does not have a direct equivalence in the semantic types and will be modeled according to the judgment of the modeler – he/she must know, for instance, that Caterpillar is a partition in the life of a Lepdopterum. The definition of the semantic type *Environment* points to the *Quantity* construct. Finally, the *Social Group* semantic type can be mapped to either the *Collective* or the (Institutional Agent) *Kind* constructs. Table I presents the results of this comparison.

TABLE I. SEMANTIC TYPES AND ONTOUML

Semantic Types	OntoUML construct
Animate	Kind
Human	Kind
Human/Kin	Role
Human/Rank	Role
Human/Social Group	Collective or Kind
Parts	Kind
Inanimate	Kind
Inanimate/Artefacts	Kind
Inanimate/Cel. & Weather	Kind
Inanimate/Environment	Quantity
Inanimate/Flora	Kind
* Varies	Phase

IV. APPROACH

Any conceptual modeling activity may be considered as a process that involves knowledge and perceptions that are inherently human and very subjective. However, it is possible to systematize modeling activities in a detailed list of activities, which can pose a guide to modelers. This work proposes, then, a conceptual modeling approach that supports modeling activities that may be already known or familiar but that, when ordered and organized as presented here, can help save time, avoid mistakes and create better models. Also, this approach has linguistics, mainly, semantics as a starting point, what is key to the actual understanding of the meaning of signs; it also focus on the interaction with UoD specialists and in the contextualized meaning of the signs present in the UoD descriptions. In order to make the analysis and understanding of the texts easier, they should be turned into a numbered list of sentences. The approach is illustrated in Figure 4 (UML Activity diagram notation); activities are described below.

A. Decompose text into simple sentences

A simple sentence is a sentence that does not have co-ordinate or subordinate clauses; they generally present the format Subject + verb + object, and none of these components can be a clause in itself. Consequently, a way to “break” complex sentences is by looking for co-ordinate and subordinate conjunctions and understanding how they relate clauses.

The aim here is to reach what Chomsky [22] defines as *kernel* sentences, i.e., the ones that belong to the *deep structure* of the language (meaning), without the transformations that lead to the *surface structure* of the language (form). Chomsky describes kernel sentences as active voice, affirmative simple sentences. Thus, it is also important to convert sentences from passive to active voice, whenever applicable, so that subjects and objects are clearly identified; if the resulting active voice sentence does not have an explicit subject, the word *someone* should be used as substitute. The list of simple sentences should also be numbered in the order they appear in the text, so that the reading of the sentences, in order, reproduces the reading of the text itself.

B. Create list of questions and doubts

As the modeler decomposes the text into simple sentences, he or she may realize that pieces of information are missing or find ambiguities that will have to be explained by the UoD specialist. The first thing to notice is the presence of the word *someone* in the subject position, derived from the transformation of a sentence into its active form – the modeler must be able to replace it with the correct agent. Another potential question arises, for instance, when two of the three components are repeated in two or more sentences; this may mean that there are synonyms in the texts and the best (or correct) sign must be identified. The modeler should write a list with all questions and doubts, in the same order as they appear in the text, and later ask all the listed questions to the domain specialist. This activity can be executed in parallel to the decomposition of the sentences described above.

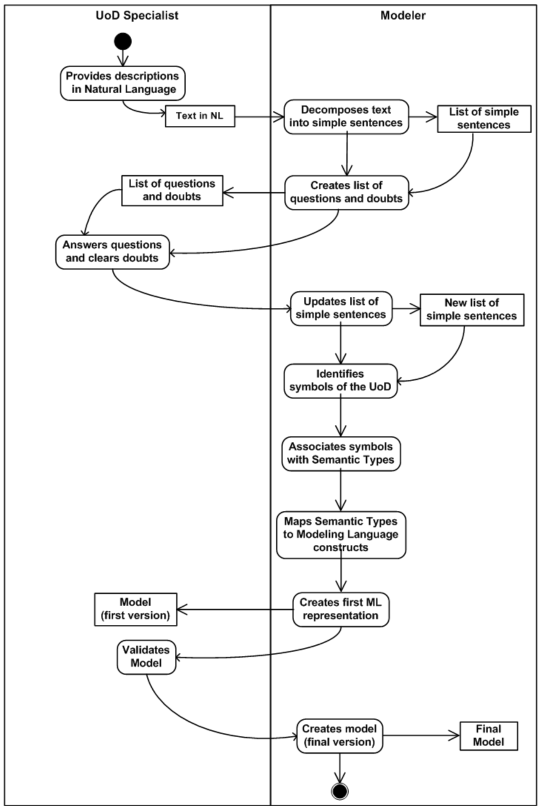
UoD Specialist
Modeler
Provides descriptions in Natural Language
Text in NL
Decomposes text into simple sentences
List of simple sentences
List of questions and doubts
Creates list of questions and doubts
Answers questions and clears doubts
Updates list of simple sentences
New list of simple sentences
Identifies symbols of the UoD
Associates symbols with Semantic Types
Maps Semantic Types to Modeling Language constructs
Model (first version)
Creates first ML representation
Validates Model
Creates model (final version)
Final Model

Figure 4. Activity diagram of the proposal

C. *Update list of simple sentences*

According to the answers provided by the UoD specialist, the modeler must update the list of simple sentences, explicating previously unknown subjects, and eliminating synonyms and ambiguities. It is advisable that this updated list be a separate document and that it does not replace the first list, so that the work rationale can be recovered at any time.

D. *Identify signs of the Universe of Discourse*

From the final list of simple sentences the modeler can identify the conceptually significant NL signs that used in the UoD being modeled. In English, as well as in other similarly structured languages, such symbols will be nouns, verbs and adjectives.

E. *Associate signs with semantic types*

The modeler must associate each of the identified signs it 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 UoD). A good practice would be to identify the nouns of each sentence and establish whether they are agents or patients; then analyze the verb that link them and identify with which semantic type such verbs should be associated. Finally, adjectives should be noted as the values that can be assigned to characteristics of the class named by the nouns. This is a very important and yet subjective activity which success rely on the modeler's knowledge, experience and ability to focus on the current context and not to allow him/herself to be influenced by his/her previous experiences. This is one of the two most important steps of the proposed approach and, for it to be performed successfully, the modeler must seek a clear understanding of concepts of the UoD being modeled.

F. *Map semantic types to OntoUML constructs*

The second most important step in the proposed approach is, no doubt, the mapping of semantic types to the ML constructs. The modeler must compare the (meta-) properties of each semantic type identified to the (meta-) properties of OntoUML constructs searching for equivalences, as it was shown above. In order to avoid repetition, the rationale and guide for this step will be explained in the example that follows.

G. *Create first version of the model*

Once the semantic types have been mapped to the ML constructs, the first version of the model can be created. Benevides and Guizzardi [8] present an OntoUML Graphical Editor that can be used for this activity. This Editor not only provides for an easier way to draw the model classes but also automatically verifies grammatical constraints pertaining to OntoUML. Such a tool ensures the syntax quality of the produced model, since it does not allow for the construction of syntactically invalid models.

H. *Validate model*

Validation should take into account language grammar and content. Recent works have also provided automated support for validation of OntoUML models via visual simulations [23]. In order to validate the content of the model, it must be understood and approved by the UoD specialist. The modeler can, once again, ask questions, this time about the (meta)properties of the constructs used to model UoD concepts, in order to make sure they were correctly mapped.

I. *Create final version of the model*

The modeler must adjust the model according to the comments provided by the UoD specialist during the validation activity. Also, there are metrics for the quality of conceptual models that, although outside the scope of the present work, must also be taken into consideration in evaluating the correctness of created model, as, for instance, the ones described in [21]. The final version of the model must take such metrics into account.

V. EXAMPLE

In their seminal conceptual modeling book [1], Battini *et* *al* provide exercise case studies for students. The NL text used in the example presented here is an excerpt that was taken from one of these case studies (pp 268--269), as follows:

"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."

According to the approach proposed in this work, a modeler should start by decomposing the text above into simple sentences and, simultaneously, write down a list of questions and doubts to be answered by UoD specialists. Table II presents the list of simple sentences that should be the result of this activity and Table III presents the question and answered produced.

TABLE II. LIST OF SIMPLE SENTENCES

No.	Sentence
0	In the library of a computer science department This "sentence" is, in fact an adverbial phrase that identifies the UoD; its signs refer to instances or individuals that, as so, will not be modeled; thus it was given the number zero.
1	Researchers can purchase books This sentence is the active voice form of the one in the NL text.
2	Students can purchase books This sentence is also the active voice form of the one in the NL text.
3	Researchers pay for books with grants. A Grant is an amount of money given by an organization or institution for a particular purpose.
4	Researchers must indicate the grant used to pay for the book Grants represent something identifiable.
5	Each student has a limited budget Budget refers to an amount of money provided for a particular purpose.
6	Each student pays for books from their budget
7	The Dean of the college fixes students' budgets every year

TABLE III. MODELER QUESTION AND ANSWER PROVIDED BY UOD SPECIALIST

No.	Question	Answer given by Specialist
1	When you say "grant", do you refer to the amount of money or to a document, like a grant report, or a grant certificate?	*It refers to the amount of money (library budget).*

In the next step, the modeler must update the list of sentences as to reflect the answers provided by the specialist. In this small case, where we had just one question, the list does not need to be updated; the modeler must just keep in mind that "grant", in this UoD, refers to the amount of money received by an institution and linked to a researcher. Next, the modeler must identify the signs in the sentences. Table IV presents the signs identified in each sentence.

TABLE IV. LIST OF SIGNS IDENTIFIED IN EACH SENTENCE

1	Researcher, Purchase, Book
2	Student, Purchase, Book
3	Researcher, Book, Grant
4	Researcher, Indicate, Grant, Book,
5	Student, Limited Budget
6	Student, Book, Budget
7	Dean, Fix, Student [yearly] Budget

The next step is the association of identified signs with semantic types. Table V presents the list of signs and its associations.

TABLE V. ASSOCIATION OF SIGNS WITH SEMANTIC TYPES

Sign	Semantic Type
Researcher	Sign refers to a person from the hierarchical distinction this person has, in relation to the other members of a group: in the group of professors some are researchers. The semantic type should be **Rank**.
Purchase	Purchase is a sign is a Primary A verb, of the type **Giving**, that, as state before, involves 3 semantic roles (from which at least two must always be present): a Donor, a Gift and a Recipient. In this case, the agent of purchase is the Recipient (here, either a Researcher or a Student) and the thing purchased if the Gift. The vendor, if present, would be the Donor.
Book	Sign refers to an object (concrete and inanimate) produced by men, thus, an **Artefact**.
Student	Sign refers to a person from the hierarchical distinction this person has, in relation to the other members of a group; 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 is an **Activity** (a semantic type associated with a noun that is derived from a verb), which relates to a Primary A verb of the type **Giving**, since an institution gives an amount of money to a Researcher; each of such occurrences must be uniquely identified.
Budget	Sign refers to an amount of money set aside for a particular purpose; a nominalization of the verb *to budget*, it is an **Activity** (a semantic type associated with a noun that is derived from a verb) which relates to a Primary A verb of the type **Giving**; the Dean is the Donor, the student is the Recipient and the Gift is the amount of money allocated to the Student; this activity is repeated every year for each Student.
Dean	Sign refers to a person from the hierarchical distinction this person has, in relation to the other members of a group; the semantic type should be

Sign	Semantic Type
	Rank.

The modeler then must map or relate semantic types to OntoUML constructs. Table VI presents construct mappings.

TABLE VI. ASSOCIATION OF SEMANTIC TYPES WITH ONTOUML CONSTRUCTS

Semantic Type	OntoUML Construct	Rationale
Rank	Role	Qualifies the entity according to a position and/or responsibility that is not exclusive (anti-rigid) but that is relationally dependent. A Role subsumes a Kind, in this case, *Person*.
Giving	Relator	A relation that is existentially dependent (moment) of at least two of three semantic roles.
Artefact	Kind	Concrete reference = existentially independent, identifiable (substance sortal); functional complex either instance of a natural kind or an artifact.

Finally, the modeler can create a first version of the model, which will be validated by the UoD specialist later. A central concept in this domain is *Purchase*. Let us for a moment, separate the two cases, i.e., purchases performed by Students (henceforth named StudentPurchase) and purchases performed by Researcher (henceforth named ResearcherPurchase). In the first case (figure 5), the stereotypes of StudentPurchase, Dean, Book, Budget and Student come directly from our analysis. Following the ontological rules underlying OntoUML, we have that roles must specialize an ultimate kind which supplies the principle of identity for its instance. In this case, the subsuming kind for both Student and Dean is Person. Person is depicted in gray in this model to highlight the fact it originates from the ontological consistency rule of OntoUML and not from our analysis using Semantic Types.

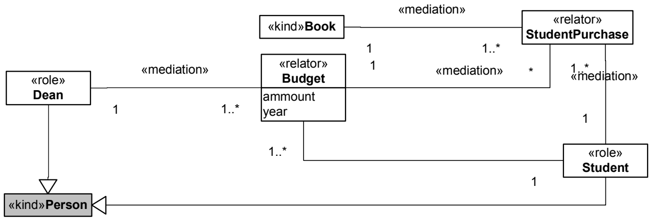

Figure 5. Student purchase - OntoUML model

In the second case, once again, the stereotypes of ResearcherPurchase, Book, Researcher and Grant come directly from our analysis; we identify Researcher as the *recipient* of the Grant and know that there is a supporting organization as the *donor*. As a Social Group, a supporting organization can be either a Kind (representing an institutional agent) or a collective. In this model, we adopt the first choice. Again, following the ontological rules of OntoUML, we identify Person as the kind subsuming Researcher. Figure 6 presents ResearcherPurchase.

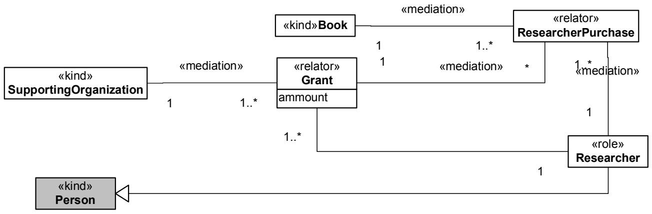

Figure 6. Researcher purchase - OntoUML model

Models presented in figures 5 and 6 can be abstracted into a single model which is illustrated two parts here (due to lack of space) in figures 7 and 8. The grey part in these figures represents the common parts of both models. Here, both types of purchase are abstracted in one single supertype (also a relator given the rules of OntoUML). A Purchase is a general relator connecting a Gift (Book) and a Recipient. The general recipient that can be either a Student or a Researcher in this case is Buyer. Since all instances of Buyer belong to same kind, then Buyer is also a sortal specializing this unique Kind (Person). The relators Budget and Grant can also be abstracted into a common supertype, since they both represent an amount of money granted by a Donor to an academic Recipient, who, in this case, is represented by Buyer. The donor, instead, is represented by the type Granter, which is a superclass of Dean and SupportingOrganization. Notice that, Granter has instances that belong to different kinds (SupportingOrganizations and Persons). Moreover, it is a semi-rigid type, i.e., a type which is rigid for some of its instances (SupportingOrganization) and anti-rigid to others (Person). Thus, following the rules of OntoUML, we have that Granter must be modeled as a **Mixin**.

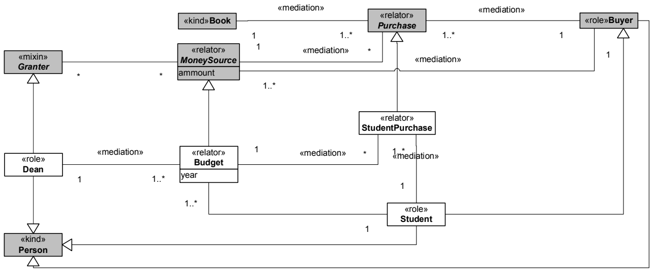

Figure 7. OntoUML model part 1

I. CONCLUSION

Conceptual modeling remains the most important activity in database design and it is not likely that this activity can ever be fully automated, since it relies on basically human-centered activities. Also, it is certain that, no matter the source of information the modeler has access to, it is always expressed in a NL. Language understanding is the ground stone for the modeling (translation) process, thus, the knowledge and application of linguistic principles are an invaluable support.

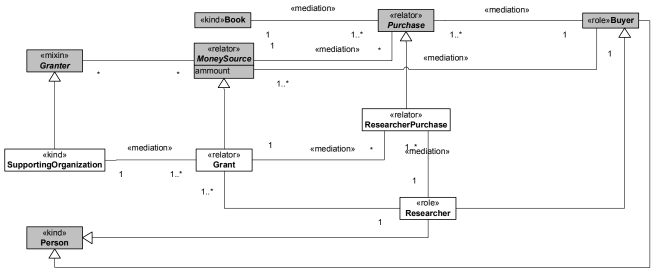

Figure 8. OntoUML model part 2

This work proposed an approach to the modeling process that is based on linguistic analysis; it relies on the semantic types theory and classification proposed by Dixon [6], and the mapping of each of those types to a well-founded ontological ML, OntoUML. The conceptual models depicted above have been produced by applying two complementary systematic procedures: first, the rules presented in this article that elaborate the mapping between semantic types to OntoUML constructs; and second, the ontological rules inherent to the OntoUML language and encoded in its metamodel, which, in this case, helped us to produce a core of most abstract classes that can be used to capture general domain constraints such as, for example, that `Purchase.MoneySource.Buyer = Purchase.Buyer`. Future work will concentrate on detailing mapping between semantic types and OntoUML constructs, and on evaluating the approach on case studies.

ACKNOWLEDGMENT

The authors thank Ms. Leila Andrade, a Librarian who kindly took the time to contribute to our case study.

This research is partially funded by the Brazilian Research Funding Agencies FAPES (Process Number 45444080/09) and CNPq (Process Number 481906/2009-6).

REFERENCES

[1] Batini, C., Ceri, S., Navathe, S.: Conceptual Database Design. Benjamin/Cummings, Redwood City (1992)

[2] Anaby-Tavor, A., Amid, D., Fisher, A., Ossher, H., Bellamy, R., Callery, M., Desmond, M., Krasikov, S., Roth, T., Simmonds, I., Vries, J.: An Empirical Study of Enterprise Conceptual Modeling. In Laender, A. H. F., Castano, S., Dayal, Umeshwar, Casati, F., Oliveira, J. P. M. (eds) ER 2009, LNCS, vol. 5829, pp 55--69, Springer, Heidelberg (2009)

[3] Gangopadhyay, A.: Conceptual Modeling from Natural Language Functional Specifications. Artificial Intelligence in Engineering, vol. 15, issue 2, 207--218 (2001)

[4] Castro, L., Baiao, F., Guizzardi, G.: A Survey on Conceptual Modeling from a Linguistic Point of View. Technical Report, Rela Te-DIA (2009)

[5] Chen, P.: English Sentence Structure and Entity-Relationship Diagrams. Information Sciences vol, 29, issues 2-3, 127-149 (1983)

[6] Dixon, R. M. W.: A Semantic Approach to English Grammar. Oxford University Press, Oxford (2005)

[7] Guizzardi, G.: Ontological Foundations for Structural Conceptual Models. CTIT, Enschede (2005)

[8] 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)

[9] 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, vol. 29, pp 224--236, Springer, Heidelberg (2009)

[10] 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)

[11] Bunge, M.: Philosophical Dictionary. Prometheus Books, Amherst (2003)

[12] McWhorter, J.: The Power of Babel: A Natural History of Language. HarperCollins, New York (2003)

[13] Saussure, F.: Course in General Linguistics. Cultrix, São Paulo (2006) In Portuguese

[14] Jackendoff, R.: Foundations of Language. Oxford University Press, Oxford (2002)

[15] Ullmann, S.: Semantics: An Introduction to the Science of Meaning. Calouste, Lisboa (1977) In Portuguese

[16] Greenbaum, S.: Oxford English Grammar. Oxford University Press, Oxford (1996)

[17] Quirk, R., Greenbaum, S.: A University Grammar of English. Longman, London (1973)

[18] Schichl, H.: Models and History of Modeling. In: Kallrath, J.: Modeling Language in Mathematical Optimization. Pp 25--36, Kluwer Academic Publishers, Norwell (2004)

[19] Guizzardi, G.: The Role of Foundational Ontologies for Conceptual Modeling and Domain Ontology Representation. In: 7th International Baltic Conference on Databases and Information Systems. Vilnius (2006)

[20] Mounin, G.: Les Problèmes Théoriques de la Traduction. Cultrix, São Paulo (1975) In Portuguese

[21] Poels, G., Nelson, J., Genero, M., Piattini, M.: Quality in Conceptual Modeling – New Research Directions. In: Olivé, A. (eds) ER 2003 Ws. LNCS, vol. 2784, pp 243--250, Springer, Heidelberg, (2003)

[21] Bechara, E.: Moderna Gramática Portuguesa. Nova Fronteira, Rio de Janeiro (2009) In Portuguese

[22] Chomsky, N.: Syntactic Structures. Mouton de Gruyter, New York (2002).

[23] Braga, B., Almeida, J., Guizzardi, G., Benevides, A.: Transforming OntoUML into Alloy: towards conceptual model validation using a lightweight formal method. In: Innovations Syst Softw Eng, Springer-Verlag, London, (2010)