

Interlinguas: A classical Approach for the Semantic Web. A Practical Case¹

Jesús Cardeñosa, Carolina Gallardo, Luis Iraola

Validation and Business Applications Research Group
Facultad de Informática. Universidad Politécnica de Madrid
28660 Madrid, Spain
{carde,carolina,luis}@opera.dia.fi.upm.es

Abstract. An efficient use of the web will imply the ability to find not only documents but also specific pieces of information according to user's query. Right now, this last possibility is not tackled by current information extraction or question answering systems, since it requires both a deeper semantic understanding of queries and contents along with deductive capabilities. In this paper, the authors propose the use of Interlinguas as a plausible approach to search and extract specific pieces of information from a document, given the semantic nature of Interlinguas and their support for deduction. More concretely, the authors describe the UNL Interlinguas from the representational point of view and illustrate its deductive capabilities by means of an example.

1 Introduction

Many activities that revolve around an advantageous use of the web are based on the efficiency to find not only documents but also on the capability to find a specific piece of information concerning a specific question from the user. The generation of a precise answer to a query requires, on the one hand, a process of semantic understanding of the query and, on the other, deductive capabilities to generate the answer.

Most recent approaches are based on the representation of contents according to the XML standard [1], where the structure of a document is explicit. XML is particularly adequate to find explicitly marked data. If those data are not explicitly marked and they are “only” deducible, current models can only be assisted by Natural Language Processing (NLP) techniques in order to find a solution to a given query. However, current NLP models generally lack of any sort of deductive capability.

In this paper, a new approach based on a classical concept in Artificial Intelligence is described. This approach is based on the use of interlinguas to represent contents, thus rescuing these representational systems from oblivion after their failure in the early nineties, when they did not meet their expectations in the Interlingua-based Machine Translation systems. The use of a concrete interlingua, Universal Network-

¹ This paper has been sponsored by the Spanish Research Council under project HUM-2005-07260 and the UPM project EXCOM-R05/11070

ing Language, will be justified, and the utility of its deductive capabilities for question answering systems will be explained by means of a short example.

2 Interlinguas

Interlinguas are mainly defined by the following characteristics:

1. The interlingual approach attempts to find a meaning representation common to many (ideally to all) natural languages, a representation that leaves aside ‘surface’ details and unveils a common structure.
2. An interlingua is just another language in the sense that it is autonomous and thus its components need to be defined: vocabulary and semantic relations mainly.
3. Senses and not words are usually the semantic atoms of interlinguas.
4. Thematic and functional relations are established among the semantic atoms of the interlingua. These relations, being semantic in nature, allow for universality and depth of abstraction and analysis.

However, although interlinguas may very well provide the knowledge representation mechanisms required both for machine translation (MT) and multilingual text generation as well as for large scale knowledge representation tasks, there are some obstacles in the design and further use of an interlingua:

- For multilingual generation and MT purposes, interlinguas are so close to the knowledge level that text generation is hindered by the lack of surface information.
- The design of an interlingua is a highly complex task; it has been proved almost unfeasible to find a suitable way to represent word meanings that is at the same time a) able to accommodate a wide variety of natural languages, b) easy to grasp and use, c) precise and unambiguous and d) expressive enough to capture the subtleties of word meanings expressed in natural languages.

The issue of representing the knowledge contained in texts written in a natural language is not new. It dates back to pioneering work in knowledge representation in the AI field [2], [3]. Interlinguas appeared after the creation of knowledge representation languages based on natural language. Developed within the MT field, classical interlinguas include ATLAS [4] or PIVOT [5]. These interlinguas are paradigmatic of the dominant approach towards interlinguas; they are designed as a general domain representational system for a large number of natural languages.

Interlingua-based MT systems did not meet the expectations they created, mainly due to the linguistic problems posed by their insufficiency to express surface phenomena and to an incomplete and unsatisfactory account of lexical meaning. However, the development of interlinguas continued and classical interlinguas evolved into the so-called Knowledge Based Machine Translation Systems. Under this label are included the KANT interlingua [6] and the Text Meaning Representations of the Mikrokosmos system [7]. These developments highlight the knowledge representation dimension of the interlingua as well as the linguistic aspects, adopting an ontological and frame-based approach for the definition of the concepts. However, the burden of such an intense and detailed knowledge based conceptual modelling can only be afforded in specific domains and for a limited number of language pairs.

Other interlingual devices such as Lexical Conceptual Structures (LCS) [8] are based on sophisticated lexical semantics analysis oriented by linguistic theories [9]. LCS representations are based on a limited number of primitive concepts that serve as building blocks for the definition of all remaining concepts. This approach is well suited for semantic inference, but at the expense of limiting the capabilities of representing the lexical richness present in natural languages.

These interlinguas are hindered by the fact that they are restricted to specific domains. Besides, they require substantial work for building up a conceptual base. The use of semantic primitives may be justified for inferential purposes but their actual design and application in a multilingual (or simply in a NLP environment) is difficult and they pose more problems than they solve. However, the use of classical interlinguas, together with similar deep semantic representations, has been reconsidered in recent years, due to the necessity of designing advanced search engines to support the Semantic Web. The use of Conceptual Graphs is an example, with some interesting results, as shown in [10], [11].

In the next section, it will be presented a new approach based on the use of an Interlingua that produces a content representation that removes away the details of the source language, so qualifying as a language independent representation.

3 The Universal Networking Language (UNL)

During the nineties, the University of the United Nations developed the Universal Networking Language (UNL), a language for the representation of contents in a language independent way, with the purpose of overcoming the linguistic barrier in Internet. It was only after years of intensive research and great efforts when the set of concepts and relations allowing the representation of any text written in any natural language was defined. This language has been proven tractable by computer systems, since it can be automatically transformed into any natural language by means of linguistic generation processes, just following its specifications [12].

The UNL is composed of three main elements: universal words, relations and attributes. Formally, a UNL expression can be view as a semantic net, whose nodes are the Universal words, linked by arcs labelled with the UNL relations. Universal Words are modified by the so-called attributes. The specifications of the language formally define the set of relations, concepts and attributes.

3.1 Universal words

They constitute the vocabulary of the language, i.e., they can be considered the lexical items of UNL. To be able to express any concept occurring in a natural language, the UNL proposes the use of English words modified by a series of semantic restrictions that eliminate the lexical ambiguity present in natural languages. When there is no English word suitable for expressing a particular concept, the UNL allows the use of words coming from other languages. Whatever the source, universal words usually require semantic restrictions for describing precisely the sense or meaning of

the base word. In this way, UNL gets an expressive richness from the natural languages but without their ambiguity. For example, the verb “land” in English has several senses and different predicate frames. Corresponding UWs for two different senses of this verb in UNL would be:

1. The plane landed at the Geneva airport.

land(icl>do, plt>surface, agt>thing, plc>thing)

This UW corresponds to the definition “To alight upon or strike a surface”. The proposed semantic restrictions stand for:

- **icl>do**: (where *icl* stands for *included*) establishes the type of action that “lands” belongs to, that is, actions initiated by an agent.
- **plt>surface**: (where *plt* stands for *place to*) expresses an inherent part of the verb meaning, namely that the final direction of the motion expressed by “land” is onto a surface.
- **agt>thing, plc>thing**: (where *agt* stands for *agent* and *plc* stands for *place*) establish the obligatory semantic participants of the predicate “land”.

2. We (agt) landed on a lonely island (plc):

land(icl>do, src>water, agt>thing, plc>thing)

This UW corresponds to the definition “To come to land or shore”. This UW differs from the previous one in the restriction *src>water* (*src* standing for *source*) that expresses an inherent part of the verb meaning, namely that the motion expressed by “land” is initiated from water. Although this method is far from perfect, it shows some advantages. Firstly, there is a consensual and “normalized” way to define UWs and how they should be interpreted. Thus, the meaning of stand-alone UWs can be easily grasped. Secondly, it is devoid of the ambiguity inherent to natural language vocabularies.

A first reproach that could be made to this interlingual vocabulary is its anglo-centred vision, which may aggravate the problem of lexical mismatches among languages. However, this system permits and guarantees expressivity and domain independency. For a more comprehensive view of the UW system, the reader is referred to [13].

The complete set of UWs composes the **UNL dictionary**. The UNL dictionary is complemented with local bilingual dictionaries, connecting UWs with headword (or lemmas) from natural languages. Local dictionaries are formed by pairs of the form:

<Headword, UW>

Where Headword is any word from a given natural language and UW the corresponding representation of one of its senses in UNL. The following are pairs linking Spanish headwords with their UWs:

1. <aterrizar, land(icl>do, plt>surface, agt>thing, plc>thing)>
2. <desembarcar, land(icl>do, src>water, agt>thing, plc>thing)>

The UNL dictionary constitutes a common lexical resource to all natural languages currently represented in the project, so that word senses of different natural languages become linked via their common UWs.

3.2 Relations

The second element of UNL is a set of conceptual relations. Relations form a closed set defined in the specifications of the interlingua that characterise a set of semantic notions applicable to most of the existing natural languages. For instance, the notion of initiator or cause of an event (its agent) is considered one of such notions since it is found in most natural languages. The current specification of UNL includes 41 conceptual relations. They are best presented grouping them into conceptually related families:

- **Causal relations:** including *condition*, *purpose*, or *reason*.
- **Temporal relations:** including instant, period, sequence, co-occurrence, initial time or final time.
- **Locative relations:** including physical place, origin, destination, virtual place, intermediate place and affected place.
- **Logical relations:** these are conjunction, disjunction, attribution, equivalence and name.
- **Numeric relations:** these are quantity, basis, proportion and range.
- **Circumstantial relations:** *method*, *instrument* and *manner*.
- **Argument relations:** agent, object, goal and source.
- **Secondary argument relations:** co-agent, co-object, co-attribution, beneficiary, and partner.
- **Nominal relations:** *possession*, *modification*, *destination*, *origin* and *meronymy (part of)*.

These relations are complemented with three additional ones, which are only used for constructing semantic restrictions for UWs, they are:

- **icl:** meaning *included in*, a hypernym of a UW.
- **equ:** meaning *equal to*, a synonym of a UW.
- **iof:** meaning *instance of*, an instance of a class denoted by an UW.

Selecting the appropriate conceptual relation plus adequate universal words allows UNL to express the propositional content of any sentence. For example, in a sentence like “The boy eats potatoes in the kitchen”, there is a main predicate (“eats”) and three arguments, two of them are instances of argumentative relations (“boy” is the *agent* of the predicate “eats”, whereas “potatoes” is the *object*) and one circumstantial relation (“kitchen” is the *physical place* where the action described in the sentence takes place).

The UNL specifications provide a definition in natural language of the intended meaning of these semantic relations and establish the contexts where relations may apply, such as the nature of the origin and final concept of the relation. For example, an agent relation can link an action (as opposed to an event or process) and a volitional agent (as opposed to a property or a substance).

3.3 Attributes

Contextual information is expressed in UNL by means of *attributes labels*. These attributes include notions such as:

- Information depending on the speaker, such as the time of the described event with respect to the moment of the utterance, the communicative goal of the utterance, epistemic or deontic modality.
- Contextual information affecting both to the participants and to the predicate of the sentence, such as aspect, number (and gender) of participants and negation defined as the “complement set” denoted by an entity.
- Pragmatic notions that affect the presentation of the information (what is considered to be the *theme* and *topic* of the sentence), reference of the entities contained in a UNL graph (UNL distinguishes between *definite*, *indefinite* and *generic* reference) and discourse structuring.
- Typographical and orthographical conventions. These include formatting attributes such as *double quotations*, *parenthesis*, *square brackets*, etc.

Attribute labels are attached to UWs and have the following syntax:

.@<attribute_label>

4 Knowledge Representation with UNL

The UNL code takes the form of a directed hyper-graph. *Universal Words* constitute the nodes of the graph, while arcs are labelled with *conceptual relations*. The graphical representation of the UNL graph corresponding to the sentence “The boy eats potatoes in the kitchen” is graphically shown in figure 1. In the graph, @def means an entity or concept with definite and known reference; @pl means plurality and @entry designate the head of the sentence.

Any UNL graph is canonically presented in textual form as a set of arcs. The syntax of each arc is as follows:

<name of the relation> (<source UW> , <target UW>)

Figure 2 displays the textual form of the UNL graph. By means of these three components UNL clearly differentiates between propositional meaning and contextual meaning of linguistic expressions: the part of the graph consisting of the universal words plus the conceptual relations represents the propositional part of a given text. The addition of UNL attributes to that graph conveys the pragmatic and contextual information of the linguistic act.

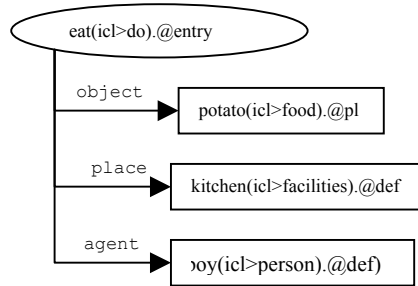


Fig. 1. Graphical representation of a UNL expression.

```

3:1]
source_sentence}
re boy eats potatoes in the
itchen
/ source_sentence}
unl}
jt(eat(icl>do).@entry,
  boy(icl>person).@def)
lc(eat(icl>do).@entry,
  kitchen(icl>facilities).@def)
oj(eat(icl>do).@entry,
  potato(icl>food).@pl)
/unl}
/S]

```

Fig 2. Textual representation of UNL

4.1 UNL for Knowledge Inference

When there is a need of representing knowledge in a domain-independent way, researchers turn back to natural language (e.g. Wordnet, the Generalized Upper Model [13] or even CyC²) to explore the “semantic atoms” that knowledge expressed in natural languages is composed of. UNL follows this philosophy, since it provides an interlingual analysis of natural language semantics. The reasons why UNL could be backed as a firm knowledge representation language can be summarised in the following points:

1. The set of necessary relations existing between concepts is already standardized. Although some of these *conceptual relations* have a strong linguistic basis (such some uses of the “obj” or “aoj” relation) other relation groups such as the logical (conjunction, disjunction), temporal, spatial and causative (condition, instrument, method) relations have been widely employed in semantic analysis as well as in knowledge representation.
2. Similarly, the set of necessary attributes that modify concepts and relations is fixed and well-defined, guaranteeing a precise definition of contextual information. Thus, UNL provides mechanisms to clear-cut propositional from contextual meaning.
3. The *semantic atoms* (UWs) are not concepts but word senses, mainly extracted from the English lexicon for convenience reasons and (implicitly) organized according to hierarchical relations, like those present in Wordnet.
4. UNL syntax and semantics are formally defined.

But to really serve as a language for knowledge representation and extraction, UNL must support deduction mechanisms. These deduction mechanisms are based on a set of semantic restrictions that implicitly make up a knowledge base (KB). This KB is endowed with classical relations, such as the *is-a* relation (represented in UNL by “icl”), synonymy (“equ” relation) and *part-of* relation (“pof” relation).

² <http://www.cyc.com>

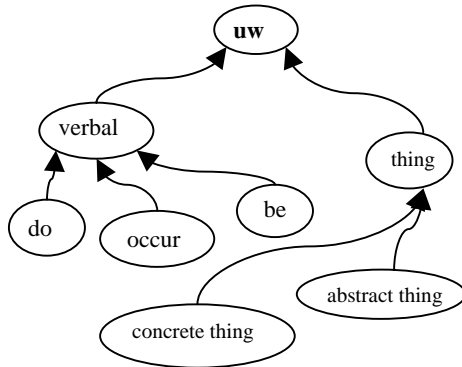


Fig 3. Upper levels of the KB

The upper levels of the KB are fixed and language neutral. Terms such as “thing” (standing for any nominal entity), “abstract thing”, “concrete thing”, “do” (verbal concepts denoting an action or an activity), “occur” (verbal concepts denoting a process) or “be” (verbal concepts denoting a state or a property) are believed to subsume all the concepts of any language. Figure 3 shows a fragment of the upper levels of the KB, where all arrows stands for the

“included in” (icl) relation.

However, as far as the terminal leaves of the hierarchy are concerned, the UNL KB adopts a maximal position, to the extent that any word sense present in any natural language is a candidate to be inserted in the KB, without any further decomposition into semantic primitives.

Non-taxonomic relations become of paramount importance in the KB, since they constitute the main mechanism for establishing the combinatory possibilities of UWs, and thus constraint the creation of coherent knowledge bases. For example, the verbal concept “do” is linked to “thing” by means of the “agent” relation (figure 4), thus imposing the obligatory presence of an agent for the verbal concept “do” and all its descendents. On the other hand, verbal concepts under “occur” are characterized by the absence of an agent; therefore an arc like the one in figure 5 would be rejected by the knowledge base.

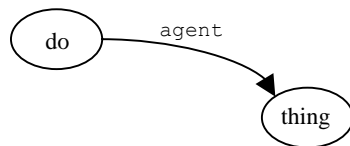


Fig 4. A non taxonomic relation

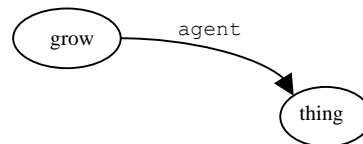


Fig 5. An incorrect relation

From an *extensional* point of view, UNL relations can be viewed as a finite set of tuples of the form $\langle \text{semantic relation}, uw_1, uw_2 \rangle$. Given the huge amount of tuples that it may contain, the UNL KB is best viewed from an *intensional* point of view as a first order logical theory composed of a finite set of axioms and inference rules. Most of the axioms state plain semantic relations among UWs, now viewed as atomic formulas of the form $\text{relation}(uw_1, uw_2)$. See some examples of the “evolution” from tuples into formulas, being “icl” and “agt” abbreviations for “included” and “agent” respectively:

$\langle \text{icl}, \text{helicopter}, \text{concrete thing} \rangle \rightarrow \text{icl}(\text{helicopter}, \text{concrete thing})$
 $\langle \text{icl}, \text{ameliorate}, \text{do} \rangle \rightarrow \text{icl}(\text{ameliorate}, \text{do})$
 $\langle \text{agt}, \text{do}, \text{thing} \rangle \rightarrow \text{agt}(\text{do}, \text{thing})$

Besides atomic formulas, the theory contains complex formulas, like the one stating the transitivity of the “icl” relation:

$$\forall w_1 \forall w_2 \forall w_3 (icl(w_1, w_2) \wedge icl(w_2, w_3) \rightarrow icl(w_1, w_3))$$

As for the inference rules, a subset of the standard rules present in first order theories may suffice for defining the relation of syntactic consequence among formulas. The UNL KB is then formally defined as the closure of the set of axioms under the consequence relation.

For any two UWs w_1, w_2 and any conceptual relation r , the UNL KB should be able to determine whether linking w_1, w_2 with r is allowed (makes sense in principle) or if it is against the intended use of w_1, w_2 and r . If the KB is viewed as a theory, the question is then if the formula $r(w_1, w_2)$ is a consequence (a theorem) of the set of axioms that form the KB or it is not. The axioms needed for answering such questions are mostly derived from the intended usage of the UNL conceptual relations and the broader semantic classes each UW belongs to.

4.2 An example of Deduction for Question Answering Systems

This section describes an example of representation that supports a question answering system. The following text deals with quite a representative building of Spain:

The edifice housing the Town Museum was designed by the architect Pedro de Ribera with the purpose of establishing an orphanage on it.

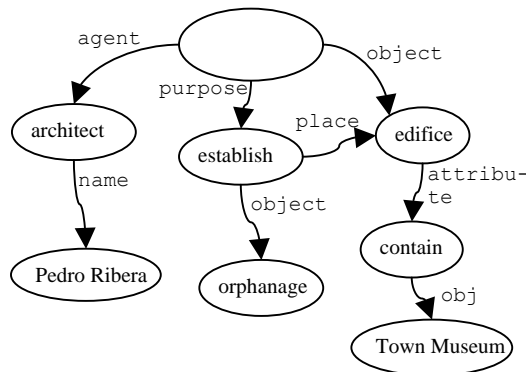


Fig 6. UNL graphical representation of text

Its UNL graphical representation is that of figure 6. If a direct query is posed to the system, the deduction process is straightforward, simply using a matching procedure for semantic nets. Let’s illustrate this procedure with the following query:

Who is the architect of the Town Museum?

This question is converted into its UNL form by means of natural language analyzing modules. Wh- questions typically

request specific pieces of information. When this query is transformed into UNL, “who” turns into the target node to be searched (that is, the unknown node), and the noun phrase “architect of the Town Museum” turns into the binary relation:

`mod(architect, “Town Museum”)`

Where “mod” simply establishes a general relation between two concepts. The next step is to link the unknown node “who” with either “architect” or “Town Museum”. The query’s linguistic structure implies that the speaker is asking for the name of a person (the entity described as architect in the question), therefore the missing relation in the query is *nam*, and the UNL representation of the query is:

```

nam(architect, ?)
mod(architect, "Town Museum")

```

In the UNL representation of the complete text, “architect” is not directly linked to “Town Museum” and in between these two nodes there is a subgraph composed of the nodes “design”, “edifice”, “contain” and finally “Town Museum” (as shown in figure 7).

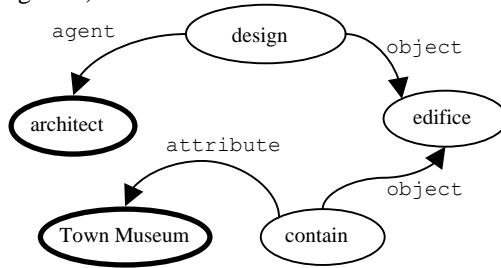


Fig. 7. Subgraph

By means of this subgraph, it can be seen that between “architect” and “Town Museum”, there exists at least one path that guarantees the connectivity between both nodes. That is, it makes sense to talk about an “architect that is related in some way to the Town Museum”³. Later, it will be searched whether there is any node pending from node “architect”

by means of the *nam* relation, which is the case. Therefore, the unknown node (the target of the query) should be “Pedro Ribera”, which is the (correct) answer to the posed query. Naturally, not always inference will be so straightforward and accurate. Many different situations may arise. But diversion does not mean here impossibility to solve problems. A model of knowledge representation based on UNL is valid for answering queries provided that the information is complete (since imprecision will not blur the inference process. That is to say, UNL can be efficiently used in closed domains where information is coherent.

5 Conclusions

Apart from multilingual text generation applications, UNL is currently being employed as text representation formalism in tasks such as information extraction, and integration with other linguistic ontologies. UNL should not be seen either as just another interlingua neither as just another knowledge representation formalism. Its goal is to serve as an intermediate *knowledge* representation that can be exploited by different knowledge intensive tasks. UNL is a formalism worth to be considered particularly in those scenarios where:

1. Multilingual acquisition and dissemination of textual information is required,
2. Deep text understanding is required for providing advanced services such as question answering, summarization, knowledge management, knowledge-based decision support, language independent document repositories, etc. For all these tasks, a domain and task dependent knowledge base is needed and building it from UNL representations presents distinct advantages over other approaches.

³ A note of caution has to be made, we are not claiming that the “mod” relation is equivalent to the combination of relations present between “architect” and “Town Museum” in the subgraph of figure 7.

3. Finally, several issues describe the UNL novelty in order to become a “de facto” standard because it is supported by a worldwide organization. On the other hand, it is necessary to remark that the representation of word senses instead of concepts increases significantly the power of UNL in comparison with other approaches. This is due to the approach to produce content representations that are closer to the linguistic surface than other representations are, thus making easier the understanding user queries.

References

- [1] <http://www.w3.org/XML/>
- [2] Quillian M.R. 1968 Semantic Memory. Semantic Information Processing. M.Minsky (Ed.), MIT press
- [3] Schank, R.C (1972). Conceptual Dependency: A Theory of Natural Language Understanding, Cognitive Psychology, Vol 3, 532-631
- [4] H. Uchida, “ATLAS-II: A machine translation system using conceptual structure as an Interlingua”, in *Proceedings of the Second Machine Translation Summit*, Tokyo, 1989.
- [5] K. Muraki, “PIVOT: Two-phase machine translation system”. In *Proceedings of the Second Machine Translation Summit*, Tokyo, 1989.
- [6] E. H. Nyberg, and T. Mitamura, “The KANT system: fast, accurate, high-quality translation in practical domains”, in *Proceedings of the 14th International Conference on Computational Linguistics (COLING '92)*, vol. 4, pp. 1254-1258, Nantes, 1992.
- [7] S. Beale, S. Nirenburg S. and G. Mahesh, "Semantic Analysis in the Mikrokosmos Machine Translation Project", in *Proceedings of the Second Symposium on Natural Language Processing (SNLP-95)*. Bangkok, Thailand. 1995.
- [8] B. Dorr, “Machine Translation Divergences: A Formal Description and Proposed Solution”, *Computational Linguistics*, vol 20(4), pp 597-633, 1994.
- [9] Jackendoff, R., *Semantic Structures*. Current Studies in Linguistics series. Cambridge, Massachusetts: The MIT Press, 1990
- [10] M. Montes, A.Gómez, A. López, A.Gelbukh. “Information retrieval with Conceptual Graph Matching”. *Lecture Notes in Computer Science*, N 1873, Springer Verlag, 2000, pp 312-321
- [11] M. Montes, A..Gómez, A..Gelbukh, A. López. “Text mining at Detail Level Using Conceptual Graphs”. *Lecture Notes in Computer Science* N 2393, Springer, 2002 pp 122-136
- [12] H. Uchida, *The Universal Networking Language Specifications*, v3.3, 2004. Available at <http://www.undl.org>.
- [13] I. Boguslavsky, J. Cardenosa, C. Gallardo, L. Iraola. “The UNL Initiative: An Overview”, *Lecture Notes in Computer Science*, vol 3406, pp 377 – 387, 2005
- [14] Bateman, J.A; Henschel, R. and Rinaldi, F. (1995) The Generalized Upper Model 2.0. <http://www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html>.