# BRIDGING THE GAP BETWEEN HUMAN LANGUAGE AND COMPUTER-ORIENTED REPRESENTATIONS

Jesús Cardeñosa, Carolina Gallardo, Eugenio Santos

*Abstract: Information can be expressed in many ways according to the different capacities of humans to perceive it. Current systems deals with multimedia, multiformat and multiplatform systems but another « multi » is still pending to guarantee global access to information, that is, multilinguality. Different languages imply different replications of the systems according to the language in question. No solutions appear to represent the bridge between the human representation (natural language) and a system-oriented representation. The United Nations University defined in 1997 a language to be the support of effective multilinguism in Internet. In this paper, we describe this language and its possible applications beyond multilingual services as the possible future standard for different language independent applications*

*Keywords: Knowledge Representation, Information modelling*

*ACM Classification Keywords: H.3.3 Information Search and Retrieval; I.2.7 Natural Language Processing:* Machine translation; *I.2.4 Knowledge Representation Formalisms and Methods:* Representation languages; *K.4.4 Electronic Commerce:* Distributed commercial transactions

## Introduction

The concept of Multimedia systems could be broadly defined as the set of systems built for transmitting information through any means that human beings are able to catch. People communicate in a natural way through the five senses, either individually or cooperatively. The evolution of technology has permitted the existence of systems in the market that reproduce the natural five senses. Among these systems and technologies, the most classical and frequent systems to transmit information utilize sight and hearing. These natural senses have permitted to see and hear. They have allowed us to visualize images of all kinds, from natural images to drawings and pictures, and from static to animated images. Further, we have been able to perceive them combined with natural or artificial sounds. However, sight and hearing do not constitute the most massive means to transmit information; the most appealed media to transmit knowledge among human beings is language, mainly recorded in its written form.

Ever since immemorial times, when human beings wanted knowledge to endure, they recorded in written texts. Even today, in the image era, images are accompanied by text. The profusion of support media for information has made that classical systems based on natural language texts incapable of organizing the information in an adequate manner, hindering a correct management of information. Several technologies like document management or information retrieval have been helped by great computational systems in order to palliate such situations.

From the second half of the XX century, when the production of written information has been massive, Internet has put in an appearance, and worldwide commerce has grown exponentially. Media companies have started to commercialize information by itself, which means that they have to conceive systems for storing information in an organized and more compact way, easier to find and thus easier to be offered to those requesting it. This requires of complex systems but also of ways of representing knowledge that permits the reliable interchange of information between humans and machines. Human language is the externalization of human knowledge and the vehicle for knowledge exchange among humans but it is inadequate for machines.

An intelligent use of knowledge for any purpose (like question answering, information retrieval, concepts deduction, etc.) calls for mechanisms of representation completely devoid of any source of ambiguity and imprecision found in natural languages, while endowed with the formal properties characteristic of computational languages.

For that, we should invest more resources in the processes of capturing, organizing and diffusing information. Thus, in the same way that we associate *knowledge* to the "sources", we should ascribe *information* to "systems", i.e., to the media in charge of providing information. That is, *knowledge* and *information* is not the same thing, and the transformation of one into the other entails something more that human language. It entails a system that permits the conversion of one into the other.

However, the design of an intermediate system between human language, as a way to represent knowledge, and systems language is not a trivial issue. In fact, human language is self-organized in an unconscious manner; it represents highly complex mental processes (be it in written or oral form) such as searching and rearranging data in a way that can are useful and comprehensible for others. On the other hand, systems for the capture and diffusion of information have a conscious organization but an information search capability quite limited. That is:

| | Knowledge Organization | Functionalities |
|---|---|---|
| Human language | Unconscious | High |
| Systems language | Conscious | Low |

The gap between the organization of human knowledge and systems knowledge still remains as the main obstacle for the construction of efficient systems with conscious knowledge, able to be easily maintained, modified or expanded. A possible approach to "fill this gap" could be an intermediate representation of knowledge (hereafter *IR*) serving as a bridge between human language and systems language. That is, a representation able to express contents with schemata and models close to human beings but at the same time a representation able to communicate with systems or represent knowledge at the systems level. Figure 1 illustrates the intermediary role that the IR language could play among human languages and systems language.

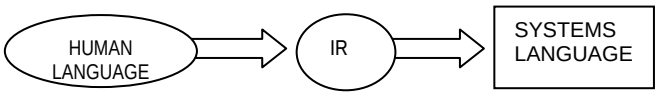


Figure 1. Intermediate Language between human language and systems language

There have already been attempts to develop these languages, although not associated to the use that we propose here. We are referring to Schank's theory of Conceptual Dependencies [Schank, 1972] to other models such as Sowa's [Sowa, 1996] that have been in fact predecessors of current ontologies. However, these models lacked a number of applications and a wider computational capability when they were proposed. In another sector of the industry, concretely language industry, we can find similar models used to represent contents in a language-independently manner. These models were known as "Interlinguas" and were used in machine translation systems such as PIVOT [Muraki, 1989] and ATLAS [Uchida, 1989].

We could say that, although with a different perspective, there are already precedents of languages and models that have attempted to overcome the gap between human languages and machine languages. We will generically call *Intermediate Representation Language* to the knowledge representation language able to make compatible human languages and systems languages.

In the late 90ies, the University of the United Nations started the development of a computational language that would allow for the representation of human knowledge in a language-independent way. The rationale behind this was to the elimination of linguistic barriers for content access in Internet. The mere fact of guaranteeing language independence involved the representation of the deepest structures of conscious human expression, so that written texts in any natural language would have the same representation in that computational language. After years of research and testing, such language has been reaffirmed as a language able to be understood by systems (for example, in order to generate any other human natural language) and it is possible to generate it just following its specifications and corresponding manuals [Uchida, 2003]).

This approach is resulting to be much more important that the initially posed (pivotal representation of texts for the elimination of linguistic barriers). This language has been gradually transformed into a knowledge representation language starting from written texts.

In this article it is presented a description of this language and how from written texts we can achieve a representation that not only "stores" knowledge but allows for an intelligent use of the resulting representation for different purposes, provided a number of functions.

## The Language

UNL is an artificial language designed for unambiguously expressing the informational content of natural language texts in a language-independent way, with the main aim of facilitating automatic multilingual information exchange on the web or in other local contexts.

Information encoded in UNL is organised into UNL documents. Since documents are commonly organised themselves into paragraphs and these into sentences, a UNL document mimics such structure and is organised into UNL paragraphs and sentences by means of HTML-like tags. UNL paragraphs consist of UNL sentences, which in turn are divided into sections. Each UNL sentence is divided into the following sections:

- The original sentence, i.e. the information that has been encoded.
- The UNL code corresponding to the original sentence.
- For each language for which a UNL Generator has been developed the automatically generated text of the UNL code into that language. Generation results are then cached in the document and available to the reader without delay. Of course, the stored results can be renewed as soon as the generators improve their output.

Besides these elements, a UNL header contains information and meta-information about the document as a whole. Although not devised in principle, UNL documents also allow for its integration into XML, since document structuring such as tables, sections, subsections, etc are not easily expressed by the UNL machinery. In fact, it is not the objective of UNL to provide document structuring but to provide a semantic structuring of contents. These ideas are further explored in [Cardeñosa, 2005] and [Hailaoui, 2005].

A UNL expression takes the form of a directed hyper-graph. Its simple nodes contain the so-called *Universal Words* and its arcs are labelled with *conceptual relations*. In addition to simple nodes, hyper-nodes are also allowed as origin or destination of arcs and consist on UNL graphs themselves. In addition to universal words and conceptual relations, *attributes* are the third ingredient of a UNL hyper-graph. Attributes may occur as labels of the universal words, modifying them in certain key aspects. These are the three building blocks of the inter-lingua and we now turn to describe each one in detail.

### Universal Words

Universal words are so called because they attempt to be universally applicable to any natural language and because their meaning is derived from the meaning of natural language words. UWs are based on the English headwords. Initially any English headword is a candidate for becoming an UW, being its meaning the meaning defined in any authoritative monolingual dictionary of English. However, in deciding on English headwords as the building blocks of the vocabulary of the interlingua, two key aspects have to be resolved:

a) Inherent ambiguity in English headwords.
b) Mismatch among lexicalized concepts in English and lexicalized concepts in other natural languages.

#### *Inherent Ambiguity in English Headwords*

Most natural language words are subject to ambiguity and polysemy. A single word in a natural language contains several senses (often related, often not), so it is fairly rare that the relation between a concept and a word is one-to-one. English vocabulary is not devoid of such ambiguity and thus a system based on English headwords as interlingual concepts should establish mechanisms for reducing such ambiguity. In order to reduce ambiguity, UWs are modified by semantic restrictions. Such semantic restrictions try to select a given sense or concept of an English Headword from the others. The most basic and simple way to achieve this is by attaching to the word a hypernym.

We will illustrate the process of defining UWs taking the following sentence as input:

```
Member States should, whenever necessary or desirable, conclude
bilateral agreements to deal with matters of common interest arising out
of the application of the present Recommendation.
```

Only content words (mainly nouns, verbs, adjectives and some prepositions and adverbs) require an UW, in this sentence, the first content word is *State*, which is a highly ambiguous headword in English.

*State* can be both a noun and a verb in English. Initially, there are there are two obvious candidates for "state" (being "icl" the abbreviation for "is included in" or the traditional *"is a"* relation) using the most general hypernym as semantic restriction:

*statet(icl>thing)* → *for the nominal senses*
*state(icl>do)* → *for the verbal senses*

However these restrictions are not very informative and there is still a great degree of ambiguity in each of these potential UWs. In order to overcome this ambiguity, more and finer semantic restrictions have to be attached to the basic UWs. For this, the possibilities are the following:

1. Use of a closer hypernym.
   a. state(icl>government)
   b. state(icl>region)
   c. state(icl>circumstance {>abstract thing})
2. Use of argument structure (for verbal senses):
   a. state( {icl>do} agt>thing, obj>thing)

*Lexical Mismatches among Languages*

This is the second problem that in fact, *every IL has to tackle*, and in the context of UNL, the solution does not come at first sight. Lexical mismatches arise when:

a) For a given concept in a given language, there is not English headword. This is a frequent case for cultural-specific terms (and other not so cultural-specific).
b) For a given concept in English, there is not an appropriate term in the target natural language. Example: En. misunderstand → Sp. entender mal
c) There is not a one-to-one relation among languages, an English headword is covered in the target language by more than one headword:
   a. Corner → Sp. esquina & rincón
   b. Marry → Ru. zhenit'sja & vyxodit' zamuzh

For these three situations, a solution must be provided. So, the solution adopted in each case is the following:

a) For the lexical gap in the English language, the specifications of the interlingua propose to include the original word as the basic UW and then semantic restrictions would be added to describe the intended meaning as much as possible, like in the Japanese term *Ikebana(icl>flower arrangement)*.
b) Lexical gaps in the target languages can be considered as a "local" problem, to be treated in the dictionaries of target languages, and not in the design of the Interlingua. For example, in the case of "misunderstand" and Spanish, it will be the task of Spanish developers' dictionary to link such a word with a complex expression such as "*entender mal*".
c) When an English headword combines the meaning of more than one headword in another language, it is possible to appeal to the semantic restrictions again in order to clarify the intended concept. In fact there are two possibilities:
   a. Ignore the difference, the Anglo-centered vocabulary here is imposed, and it will be the task of local generator to choose (in the case of Russian) into the verb *vyxodit' zamuzh* or the verb *zhenit'sja*.
   b. Express the difference, with the use of the semantic restrictions like for example:
      i. *[vyxodit' zamuzh] marry(agt>female); [zhenit'sja] marry(agt>male)*
      ii. *[esquina] corner(mod>outside angle); [rincón] corner(mod>inside)*

Of course it will be desirable that all these UWs conforms a hierarchy of inter-related UWs, so that *marry(agt>female)* and *marry(agt>male)* depend of a more general UW like *marry( agt>person, obj>person)*. Thus, semantic restrictions impose themselves a hierarchy into the system of UWs. The result is the so-called UNL-KB organizing the UW concepts *à la Wordnet* [Fellbaum, 1998], thus implicitly linking and relating the vocabulary of natural languages through the pivotal UW system.

## UNL Relations

The second ingredient of UNL is a set of conceptual relations. Relations form a closed set defined in the specifications of the inter-lingua. The rationale behind conceptual relations is twofold:

1. To characterize a set of semantic notions applicable to most of the existing natural languages. For instance, the notion of initiator or cause of an event (agent) is considered one of such notions since it is found in most languages.
2. To select a small set of semantic notions relevant to produce an inter-lingual semantic analysis. The notion of agent is regarded as one of such relations, because of its central role in the analysis of the meaning of many sentences.

Therefore, a UNL representation is mainly based on a role-based description of an event or situation, following the tradition started by Fillmore's case grammars, rather than a more logical semantic analysis of sentences. When defining the intended meaning of each conceptual relation, the specifications of the language [Uchida, 2003] rely on two intensional expedients:

1. Setting semantic constraints over the universal words that can appear as first (origin) and second argument (destination) of the relation. These constraints are based on the lexical relations established among universal words in the Knowledge Base. In the case of the agent relation, such restrictions include that the origin Universal Word must denote an event that accepts an initiator (and not an event that "just happens") and the destination universal word must denote an entity (as opposed to a quality or an event, let us say).
2. Giving a natural language explanation of the intended meaning of the relation. For the agent relation this explanation may just say that the agent is the cause or initiator of the event, an entity without which the event would not happen.

Provided that the Knowledge Base is built upon unambiguously defined principles, the first mechanism gives us a rigorous characterization of the conceptual relations. The second expedient is subject to the ambiguities of natural language, and the resulting definition is therefore semi-formal.

The current specification of UNL includes 41 conceptual relations, including causal, temporal, logical, numeric, circumstantial and argument relations.

Selecting the appropriate conceptual relation plus adequate universal words allows UNL to express the propositional content of any sentence.

In the input sentence, some of these relations are exemplified. This sentence describes a main event, denoted by the English predicate *conclude* and its dependent participants. The UW for the main predicate of the sentence is *conclude(agt>thing, obj>agreement)*. It requires two participants:

a) The agent or initiator of the event: In this sentence, the agent is "Member States" that coincides with the subject of the clause.
b) An object (or theme affected by the event, required for the completion of the event), realized in the "bilateral agreement" as the direct object.

Given the syntax of UNL and the binary nature of relations, when specifying the UNL representation of the sole event "Member States should conclude bilateral agreements" the *agt* relation links *conclude(agt>thing, obj>agreement)* as source UW and *state(icl>government)* as target UW. Analogously, there is a modifying *mod* relation between *state* and *member*; and *agreement* and *bilateral.*

Figure 2 shows a graphical version of the UNL representation of the main clause of the sentence.

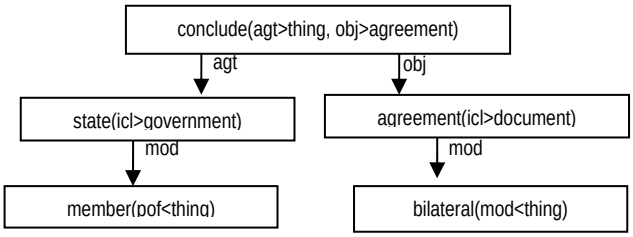


Fig. 2. Representation of the main event of the sentence

### Description of Attributes

Contextual information such as the time of the event with respect to the time of the utterance, informative structure of the sentence, speaker's communicative goal and attitudes, etc. is expressed in UNL by means of *attribute labels*. UNL attributes include notions such as:

- Information depending on the speaker, such as time of the described event with respect to the moment of the utterance; the communicative goal of the utterance; epistemic and deontic modality.
- Contextual information affecting both the participants both the predicate of the sentence, like aspectual properties of the event, number of nominal concepts.
- Pragmatic notions like the organization of the information in the original sentence, referential status of referring expressions and other labels determining discourse structure.
- Typographical and orthographical conventions. These include formatting attributes such as *double quotations, parenthesis, square brackets*, etc.

## Applications

The UNL System has an indubitable application to all the existing information systems. The introduction of multilinguality in any other system is almost a model case of added value services. However, it does not mean that multilinguality be only translation services. We will describe some possible applications of the UNL system.

### UNL as Language for Knowledge Representation

UNL is mainly used as a support language for multilingual generation of contents coming from different languages. However, its design allows for non-language centered applications, that is, UNL could serve as a support for knowledge representation in generic domains. When there is a need to construct domain-independent ontologies, researches turn back to natural language (such as Wordnet, GUM [Bateman, 1995] or even CyC [www.cyc.com]) 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 are:

1. The set of necessary relations existing between concepts is already standardized and well defined.
2. It is the product of intensive research on the thematic roles existing in natural languages by a number of experts in the area of Machine Translation and Artificial Intelligence, guaranteeing wide coverage of all contents expressed in any natural language.
3. Similarly, the set of necessary attributes that modify concepts and relations is fixed and well defined, guaranteeing a precise definition of contextual information.
4. UNL syntax and semantics are formally defined.

However, to really serve as a language for knowledge representation, it must support deduction mechanisms and must specify how a knowledge base could be build up in the UNL language. This idea is explored in [Cardeñosa, 2004].

Being a language suitable for knowledge representation, UNL could be the support of ontologies or of Cross Language Information Retrieval systems. UNL could be a firm candidate for this because of its long history as an interlingua, and the existence of analysis and generation systems to and from UNL.

### Cross-lingual Information Retrieval

To support cross-lingual information search could be one of the most appealing applications. Because the information existing in UNL is in fact independent of the original language, placing information in a web site written in UNL supposes that is accessible from any other language. But also, the search systems could try to find information based on concepts (much more effective than based on terms or keywords that are dependent on the language) and find it (if it exists) independently of the language used by the generator of the information searched. UNL offers a promising approach to this kind of systems because the search of information based on concepts is not difficult to be re-written in UNL (they could be UW) almost automatically, always under the supervision of the searcher.

## Multilingual Information System

One possible situation is that an organization has public information that should be shared and distributed in multilingual form. This is not exactly a problem of translators (at an acceptable cost) but a problem of maintenance. In this case, an organism, such as any derived from the United Nations for instance (or any other as the European Commission, Health Care Organisation, etc.) should maintain an on-line system with all kind of information about organizations. This could be classified as a simple multilingual service, where the information should be written in UNL (in the UNL Document Base) and shown through the Web in different languages. The maintenance is carried out by making changes in the UNL code. The style in which organisations are described makes post-editions unnecessary most of the times.

An additional use of this kind of system is the maintenance of technical documentary databases with multilingual necessities. One case would be the technical documentation (maintenance of industrial equipment for instance) that has to be managed in many countries, or in the case of companies with branches in several countries where a clear and precise documentation is essential for reaching a leading market position. Writing this technical documentation in UNL would clearly permit unified contents so that no differences derived from different translators could cause technical problems. It is well known that the manuals for some languages are almost in all cases translated from the original language into English and from English to the target language introducing in some cases a double risk of mistakes. The use of the UNL system for this kind of application permits also that the post-edition (if needed) can be done directly by the final users.

The multilingual Access to Public Sector Information is a general goal of big public organizations. One of the major problems to reach the objective to make the public information really available is the multilingual origin of the documents. In fact one of the recommended actions mainly to the European Industry is to affiliate contents to the Multilingual e-Content Europe portal [Nicholas, 2000]. Of course, having multilingual contents, the industry and public organizations have also to guarantee the multilingual access.

## Multilingual Transactional System

Different issues have been solved in the last years to facilitate the Business to Customer (B2C) services as the typical practice of Electronic Commerce. Most of the systems are based on the use of English language. The incorporation of multilingual capacities in companies supposes an effective increasing of market and also of image. The advantage is that the amount of work to encode any text that belongs to a web-site is the same disregarding the number of source languages. It is only needed the target language generator (at the moment there exist more than ten languages generators covering the 85% of the human population). The integration of the systems is very easy and has not special complications.

However, where the UNL system should have more impact is on the B2B activities. All the concepts and components from the B2B, but particularly the ontologies based on cXML [Merkov, 1999], to define business documents defining technical and business dictionaries (completely compatible with the UW dictionaries). OBI's [OBI, 1999] data formats rely on EDI standards for document exchanges etc. which are completely compatible with the structure of the UNL Documents. The international commercial exchange and current growth of the E-commerce are due to this kind of exchanges. Here the availability of multilingual systems able to support the exchange of documents, transactional information, a correct common understanding of contractual documentation (well addressed by a common UNL codification) is perhaps the most important application of this system. In addition, corporate information and even more complex systems (as multilingual e-mail) can be supported.

## From Bilingual to Multilingual Translation Systems

The UNL system could be viewed as an alternative to the classical machine translations systems. However, it is not exactly the case. When the classical machine translation systems massively follow the model of "transfer", the UNL is conceived in a different way. First of all, the UNL system is not a system to support machine translation but multilingual services. It is not the same. There is not any automatic conversion from a language to UNL.

For instance, analysers and dictionaries of a particular language can be integrated with the production of UNL code at the required level. An existing dictionary can be reused to develop the UW and thus to develop the UNL Dictionary for a language or a specific domain. The target language generators of an existing language can

be reused once integrated the input with the UNL code. These operations permit the transformation of a bilingual system into a multilingual one. In fact, the Russian Language Centre* [http://www.unl.ru], the French Language Centre and some others are sustaining the UNL system reusing bilingual pre-existing machine translation systems. Thus, this means that there are two types of users of this system, the industry itself manufacturing the integration and therefore creating multilingual systems with a high degree of reuse of the linguistic repository and the final users of the machine translation systems like human translators, that are normally in charge of the post-editing of the target documents. The main advantage is that these persons will increase their productivity because while working in just one language, they are producing contents in many other languages.

## References

[Bateman, 1995] Bateman, J.A; Henschel, R. and Rinaldi, F. The Generalized Upper Model 2.0. 1995. Available on line at http:// www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html

[Cardeñosa, 2004] Cardeñosa, J., Gallardo, C., and Iraola. "The forgotten key point for assuring knowledge consistency in CLIR systems". In: Proceedings of the Workshop Lessons Learned from Evaluation: Towards Transparency & Integration in Cross-Lingual Information Retrieval (LREC, 2004), Lisbon, May, 2004.

[Cardeñosa, 2005]. Cardeñosa, J., Gallardo, C., and Iraola, L. An XML-UNL Model for Knowledge-Based Annotation. Research on Computing Science, vol 12, pp 300-308. ISBN: 970-36-0226-6. México, 2005

[CyC] www.cyc.com

[Fellbaum, 1998] Fellbaum, C., editor. WordNet: An Electronic Lexical Database. Language, Speech, and Communication Series. MIT Press, 1998

[Hailaoui, 2005]. Hailaoui, N. and Boitet, C. A Pivot XML-Based Architecture for Multilingual, Multiversion Documents: parallel monolingual documents aligned through a central correspondence descriptor and possible use of UNL. Research on Computing Science, vol 12, pp 309-326. ISBN: 970-36-0226-6. México, 2005

[Merkov, 1999] Merkov, M. cXML: A new Taxonomy for E.-Commerce. 1999. Available on line at http://ecommerce.internet.com/outlook/article/0,1467,7761_124921,00.html

[Muraki, 1989] Muraki, K. PIVOT: Two-phase machine translation system. Proceedings of the Second Machine Translation Summit, Tokyo, 1989

[Nicholas, 2000] Nicholas, L.; and Lockwood, R. Final Report: SPICE-PREPII: Export potential and linguistic customization of digital products and services. EPS Ltd and Equipe Consortium Ltd, 2000

[OBI, 1999] The Open Buying on the Internet (OBI) Consortium. 1999. Open buying on the Internet. OBI specification. 2000. Available on line at http://www.openbuy.org/

[Schank, 1972] Schank, R.C. Conceptual Dependency: A Theory of Natural Language Understanding, Cognitive Psychology, Vol 3, 532-631, 1972

[Sowa, 1996] Sowa, J.F. "Processes and participants," in Eklund et al., eds. (1996) Conceptual Structures: Knowledge Representation as Interlingua, Lecture Notes in AI -1115, Springer-Verlag, Berlin, pp. 1-22, 1996

[Uchida, 1989] Uchida, H. ATLAS-II: A machine translation system using conceptual structure as an Interlingua. Proceedings of the Second Machine Translation Summit, Toky, 1989

[Uchida, 2003] Uchida, H. The Universal Networking Language. Specifications. 2003. Available on-line at http://www.undl.org

## Authors' Information

Jesús Cardeñosa – UPM.Artificial Intelligence Department, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n. 28660 Madrid, SPAIN; email: carde@opera.dia.fi.upm.es

Carolina Gallardo – UPM. Artificial Intelligence Department, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n. 28660 Madrid, SPAIN; email: carolina@opera.dia.fi.upm.es

Eugenio Santos – UPM. Business organisation Department. Universidad Politécnica de Madrid. Carretera de Valencia Km.7. 28041 Madrid, SPAIN; email: esantos@eui.upm.es

* Language Centres are the operational national units that support the development of specific local language within the worldwide organization of the UNL Programme.