# A rationale for using UNL as an Interlingua and more in various domains

**Christian BOITET**
GETA, CLIPS, IMAG
385, av. de la bibliothèque, BP 53
F-38041 Grenoble cedex 9, France
Christian.Boitet@imag.fr

*LREC-02 First International Workshop on UNL, other Interlinguas and their Applications, 1 June 2002*

**Abstract**

The UNL *language* of semantic graphs may be called as a "semantico-linguistic" interlingua. As a successor of the technically and commercially successful ATLAS-II and PIVOT interlinguas, its potential to support various kinds of text MT is certain, even if some improvements would be welcome, as always. It is also a strong candidate to be used in spoken dialogue translation systems when the utterances to be handled are not only task-oriented and of limited variety, but become more free and truly spontaneous. Finally, although it is not a true representation language such as KRL and its frame-based and logic-based successors, and although its associated "knowledge base" is not a true ontology, but rather a kind of immense thesaurus of (interlingual) sets of word senses, it seems particularly weel suited to the processing of multilingual information in natural language (information retrieval, abstracting, gisting, etc.).

The UNL *format* of multilingual documents aligned at the level of utterances is currenly embedded in html (call it UNL-html), and used by various tools such as the UNL viewer. By using a simple transformation, one obtains the UNL-xml format, and profit from all tools currently developed around XML. In this context, UNL may find another application in the localization of multilingual textual resources of software packages (messages, menu items, help files, and examples of use in multilingual dictionaries.)

**Keywords**: UNL, multilingual communication, cross-lingual information retrieval, localization

## Introduction

UNL is the name of a project, of a meaning representation language, and of a format for "perfectly aligned" multilingual documents. There is some hefty controversy about the use of the UNL language as an "interlingua", be it for translation or for other applications such as cross-lingual information retrieval. On the other hand, there is almost no discussion on the UNL format, in its current form, embedded in HTML, or some directly derivable form, embedded in XML.

We argue that the UNL language is indeed a good interlingua for automated translation, ranging from fully automatic MT to interactive MT of several kinds through, we believe, spoken translation of non task-oriented dialogues. It is also more than that, due to the associated "knowledge base", and has a great potential in textual information processing applications.

We will first give our view of what the UNL language is, and then develop a "rationale" for using the UNL language UNL along the previous lines. We will then describe some interesting potential uses of the UNL format in an "XML-ized" form.

## 1. The UNL language

The UNL representation is made of "semantic graphs" where a graph expresses the meaning of some natural language utterance. Nodes contain lexical units and attributes, arcs bear semantic relations. Connex subgraphs may be defined as "scopes", so that a UNL graph may be a hypergraph.

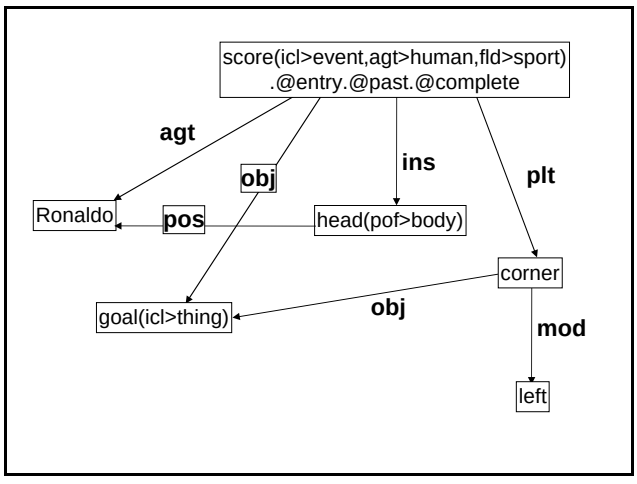


*Fig. 1: a possible UNL graph for "Ronaldo has headed the ball into the left corner of the goal"*

The lexical units, called Universal Words (in French, not "mot universel" but better "Unité de Vocabulaire Virtuel" or UVV or UW), represent word meanings, something less ambitious than concepts. Their denotations are built to be intuitively understood by developers knowing English, that is, by all developers in NLP. A UW is an English term or pseudo-term possibly completed by semantic restrictions.

A UW such as "process" represents all word meanings of that lemma, seen as citation form

(verb or noun here). The UW "process(icl>do, agt>person)" covers the verbal meanings of processing, working on, etc.

The attributes are the (semantic) number, genre, time, aspect, modality, etc.

The 40 or so semantic relations are traditional "deep cases" such as agent, (deep) object, location, goal, time, etc.

One way of looking at a UNL graph corresponding to an utterance U-L in language L is to say that it represents the abstract structure of an equivalent English utterance U-E as "seen from L", meaning that semantic attributes not necessarily expressed in L may be absent (e.g., aspect coming from French, determination or number coming from Japanese, etc.).

## 2. Some arguments for using the UNL language in various contexts

To show that using UNL is not only a workable but a good or perhaps the best idea at the moment, we can say that

- the "pivot" technique HAS BEEN not only experimented but deployed successfully (ATLAS, PIVOT, ULTRA, KANT).
- in particular, ATLAS-II (Fujitsu) is built on the basis of a pivot from which the UNL representation has evolved. The main designer of UNL, H. Uchida, was also the main designer of ATLAS-II.
- ATLAS-II has been recognized as the best EJ/JE MT system in Japan for over 10 years and has a very large coverage (586,000 words in English and Japanese).
- interlingual representations can not in principle be used (alone) to achieve the highest quality achievable by transfer systems, BUT they can give quite high quality as demonstrated by ATLAS-II.
- due to the precise nature of UNL, it is possible for human non-specialists to improve a UNL representation interactively, a posteriori, from any UNL-related language, and on demand (meaning partially — think of "lazy improvement").
- in many contexts other than translation, an interlingual, semantic-oriented representation like UNL is actually the best solution. For example, all applications related to information processing in multilingual contexts don't need a very precise representation of the FORM of the information, they need a precise ENOUGH representation of the INFORMATION CONTENT of the information.
- applications such as information retrieval and abstracting have already been prototyped successfully with UNL. It is far easier to generate SQL or SQL-like queries and answers from a UNL form than from text in many languages.

## 3. Applications of the UNL format

The UNL *format* of multilingual documents aligned at the level of utterances is currenly embedded in html (call it UNL-html). A sentence is represented between the [S] and [/S] tags. Its original text is contained between {org:el} (English, here) and {/org}, its UNL graph between {unl} and {/unl}, each French version between {fr} and {/fr}, and analogously for other languages. Atrtibutes such as version, date, location, author, etc. may appear in the tags. Here is a slightly simplified example of a file in UNL-html format.

```
<HTML><HEAD><TITLE>
Example 1  El/UNL
</TITLE></HEAD><BODY>
[D:dn=Mar Example 1, on= UNL French, mid=First.Author@here.com]
[P]
[S:1]
{org:el}I ran in the park yesterday.{/org}
{unl}
agt(run(icl>do).@entry.@past,i(icl>person))
plc(run(icl>do).@entry.@past,park(icl>place).@def)
tim(run(icl>do).@entry.@past,yesterday)
{/unl}
{cn dtime=20020130-2030, deco=man}
我昨天在公園裡跑步 {/cn}
{de dtime=20020130-2035, deco=man}
Ich lief gestern im Park. {/de}
{es dtime=20020130-2031, deco=UNL-SP}
Yo corri ayer en el parque.{/es}
{fr dtime=20020131-0805, deco=UNL-FR}
J’ai couru dans le parc hier. {/fr}[/S]
[S:2]
{org:el}My dog barked at me.{/org}
{unl}
agt(bark(icl>do).@entry.@past,dog(icl>animal))
gol(bark(icl>do).@entry.@past,i(icl>person))
pos(dog(icl>animal),i(icl>person))
{/unl}{de dtime=20020130-2036, deco=man}
Mein Hund bellte zu mir.{/de}
{fr dtime=20020131-0806, deco=UNL-FR}
Mon chien aboya pour moi. [/S] [/P][/D]
</BODY></HTML>
```

The French versions have been produced automatically while the German and Chinese versions have been translated manually.

The output of the UNL viewer for French is:

```
<HTML><HEAD><TITLE>
Example 1  El/UNL
</TITLE></HEAD><BODY>
J’ai couru dans le parc hier.
Mon chien aboya pour moi.
</BODY></HTML>
```

and will probably be displayed by a browser as:

**Example 1 El/UNL**

J’ai couru dans le parc hier. Mon chien aboya pour moi.

and similarly for all other languages.

The UNL viewer produces on demand as many html files as languages selected and sends them to any available browser.

The UNL-html format predates XML, hence the special tags like [S] and {unl}, but it is easy to derive from it an XML format and to transform the documents into an equivalent "UNL-xml" format. Then, using DOM and javaScript, it is possible to produce various views, including that of a classical viewer, a bilingual or multilingual editable presentation, and a revision interface where not only the text but the UNL graph and possibly other structures may be directly manipulated.

Let us take an example from an experiment performed for the "Forum Barcelona 2004" on documents in Spanish, Italian, Russian, French and Hindi. Hindi and Russian are not shown, but Japanese has been added by hand. The XML form is simplified.

Correct sentences are produced by the deconverters from correct and complete UNL graphs.

Suppose for the sake of illustration that some UNL graph has been produced from a Chinese version, and does not contain definiteness and aspectual information. All results may be wrong wrt articles, and some wrt aspect.

```
<unl:S num="1">
<unl:org lg="cn">在博覽會之後，城市 將獲得一片海岸域 </unl:org>
<unl:unl>
<unl:arc> agt(retrieve(icl>do).@entry.@future, city) </unl:arc>
<unl:arc> tim(retrieve(icl>do).@entry.@future, after) </unl:arc>
<unl:arc> obj(after, Forum) </unl:arc>
<unl:arc> obj(retrieve(icl>do).@entry.@future, zone(icl>place).@indef) </unl:arc>
<unl:arc> mod(zone(icl>place).@indef, coastal) </unl:arc> </unl:unl>
<unl:cn> 在博覽會之後，城市 將獲得一片海岸域 </unl:cn>
<unl:el> After a Forum, a city will retrieve a coastal zone.</unl:el>
<unl:es> Ciudad recobrará una zona de costal después Foro. </unl:es>
<unl:fr> Une cité retrouvera une zone côtière après un forum. </unl:fr>
<unl:it> Città ricuperarà une zona costiera dopo Forum. </unl:it>
<unl:jp> フォーラムの後で, 都市は沿岸水域を取り出す。  </unl:jp>
</unl:S>
```

The idea of "coedition" is applicable if there is a UNL graph associated with a segment one wants to modify. The goal is to share the revisions across languages, by reflecting them on the UNL graph, e.g.

- add ".@def" on the nodes containing "city", "Forum".
- replace "retrieve" by "recover" and add ".@complete" on the node containing it.

It is not possible in principle to deduce the modification on the graph from a modification on the text. For example, replacing "un" ("a") by "le" ("the") does not entail that the following noun is determined (.@def), because it can also be generic ("il aime la montagne" = "he likes mountains"). Hence, the technique envisaged is that:

- revision is not done by modifying directly the text, but by using a menu system,
- the menu items have a "language side" and a hidden "UNL side",
- when a menu item is chosen, only the graph is transformed, and the action to be done on the text is stored and shown next to its focus in the "To Do" zone,
- at any time, the new graph may be sent to the L0 deconverter and the result shown. If is is satisfactory, that shows that errors were due to the graph and not to the deconverter, and the graph may be sent to deconverters in other languages. Versions in some other languages known by the user may be displayed, so that improvement sharing is visible and encouraging.

New versions will be added with appropriate tags and attributes in the original multilingual document in UNL-xml format, or in a DBMS, so that nothing is ever lost, and cooperative working on a document is feasible. UNL may find another application in the localization of multilingual textual resources of software packages (messages, menu items, help files, and examples of use in multilingual dictionaries.)

Apart of the "coedition", there are many other portential applications of UNL, such as:

- crosslingual information retrieval, on which we are currently working,
- abstracting & gisting, which has been prototyped at NecTec and in India,
- localization of software packages: messages in multiple languages could be created from UNL graphs produced from a graphical interface or by enconversion, and then sent to appropriate deconverters.

For this last point, we have found how to represent messages including variables (such as integers, file names etc.), but not yet how to handle messages including morphological or even lexical variants (as "4 goda / 5 let" for "4 years / 5 years" in Russian).

## Conclusion

The UNL language is an artificial interlingua, embeddable in html or xml formats for multilingual document representation and processing. Because of its both abstract and linguistic nature, the UNL language offers many more interesting potential applications than other types of interlingua such as task and/or domain specific interlingua.

The history of MT shows that UNL will also be usable in the context of high-quality MT,

quality being obtained through typology specialization and/or interactive improvement, a priori (interactive disambiguation after all-path robust analysis) and/or a posteriori by coedition of the text in any language and the corresponding UNL graph.

## References


Blanc É. & Guillaume P. (1997) *Developing MT lingware through Internet : ARIANE and the CASH interface*. Proc. of Pacific Association for Computational Linguistics 1997 Conference (PACLING'97), Ohme, Japon, 2-5 September 1997, 1/1, pp. 15-22.

Blanchon H. (1994) *Perspectives of DBMT for monolingual authors on the basis of LIDIA-1, an implemented mockup*. Proc. of 15th International Conference on Computational Linguistics, COLING-94, 5-9 Aug. 1994, 1/2, pp. 115—119.

Boitet C., Guillaume P. & Quézel-Ambrunaz M. (1982) *ARIANE-78, an integrated environment for automated translation and human revision*. Proc. of COLING-82, Prague, July 1982, North-Holland, Ling. series 47, pp. 19—27.

Boitet C. (1994) *Dialogue-Based MT and self-explaining documents as an alternative to MAHT and MT of controlled languages*. Proc. of Machine Translation 10 Years On, 11-14 Nov. 1994, Cranfield University Press, pp. 22.21—29.

Boitet C. & Blanchon H. (1994) *Multilingual Dialogue-Based MT for Monolingual Authors: the LIDIA Project and a First Mockup*. Machine Translation, Vol. 9, N° 2, pp. 99—132.

Boitet C. (1997) *GETA's MT methodology and its current development towards personal networking communication and speech translation in the context of the UNL and C-STAR projects*. Proc. of PACLING-97, Ohme, 2-5 September 1997, Meisei University, pp. 23-57.

Boitet C., Réd. (1982) *"DSE-1"— Le point sur ARIANE-78 début 1982*. Contrat ADI/CAP-Sogeti/Champollion (3 vol.), GETA, Grenoble, janvier 1982, 616 p.

Brown R. D. (1989) *Augmentation*. (Machine Translation), Vol., N° 4, pp. 1299-1347.

Ducrot J.-M. (1982) *TITUS IV*. In *Information research in Europe. Proc. of the EURIM 5 conf. (Versailles)*, edited by Taylor P. J., London, ASLIB.

Kay M. (1973) *The MIND system*. In *Courant Computer Science Symposium 8: Natural Language Processing*, edited by Rustin R., New York, Algorithmics Press, Inc., pp. 155-188.

Lafourcade M. (2001) *Lexical sorting and lexical transfer by conceptual vectors*. Proc. of MMA'01, 29-31/1/01, SigMatics & NII, Tokyo, 10 p.

Lafourcade M. & Prince V. (2001) *Synonymies et vecteurs conceptuels*. Proc.,, 29-31/1/01, SigMatics & NII, Tokyo, 10 p.

Maruyama H., Watanabe H. & Ogino S. (1990) *An Interactive Japanese Parser for Machine Translation*. Proc. of COLING-90, 20-25 août 1990, ACL, 2/3, pp. 257-262.

Melby A. K., Smith M. R. & Peterson J. (1980) *ITS : An Interactive Translation System*. Proc. of COLING-80, Tokyo, 30/9-4/10/80, pp. 424—429.

Moneimne W. (1989) *TAO vers l'arabe. Spécification d'une génération standard de l'arabe. Réalisation d'un prototype anglais-arabe à partir d'un analyseur existant*. Nouvelle thèse, UJF.

Nirenburg S. & al. (1989) *KBMT-89 Project Report*., Center for Machine Translation, Carnegie Mellon University, Pittsburg, April 1989.

Nyberg E. H. & Mitamura T. (1992) *The KANT system: Fast, Accurate, High-Quality Translation in Practical Domains*. Proc. of COLING-92, 23-28 July 92, ACL, 3/4, pp. 1069—1073.

Sérasset G. & Boitet C. (2000) *On UNL as the future "html of the linguistic content" & the reuse of existing NLP components in UNL-related applications with the example of a UNL-French deconverter*. Proc. of COLING-2000, Saarbrücken, 31/7—3/8/2000, ACL, 7 p.

Slocum J. (1984) *METAL: the LRC Machine Translation system*. In *Machine Translation today: the state of the art (Proc. third Lugano Tutorial, 2–7 April 1984)*, edited by King M., Edinburgh University Press (1987).

Wehrli E. (1992) *The IPS System*. Proc. of COLING-92, 23-28 July 1992, 3/4, pp. 870-874.