

The UNL System

Jesús Cardeñosa, Edmundo Tovar, Carolina Gallardo

Validation and Business Applications Research Group
Artificial Intelligence Department
Universidad Politécnica de Madrid.(Spain)
E-mail: carde@fi.upm.es

Abstract

The UNL System was conceived to support multilingual services in Internet being an alternative to the classical machine translation systems. The UNL System is based in the creation of documents written in an unique computational language able to represent concepts and their relations. The definition of this language has been possible thanks to the collaboration of more than one hundred people, prestigious researchers, and scientists of all around the world. The purpose of the UNL is to break the linguistic barriers in Internet that avoid the real global access to the knowledge and culture for all people. Since the starting of the UNL project in 1996, the participants in the project from initially 14 languages have made substantial progress in the technical as well as the organizational aspects involved. This demonstration will show to the academic and business communities the current state and practical achievements of the UNL system.

1. Background

The UNL project was launched in November 1996 by the Institute of Advanced Studies, United Nations University, and its development during the next four years might be summarized as follows:

- ?? **1997:** Setting up of the pioneer teams of the project. During the first year up to fourteen teams started to work under the co-ordination of the UNL Center created for this goal. The initial languages represented included Arabic, Chinese, German, French, Japanese, Hindi, Indonesian, Italian, Mongolian, Portuguese, Russian, Spanish and Thai (later on Latvian and Korean were also included). The first version of the specifications of the interlingua was produced, as well as the first version of the tool for text generation from UNL representations generically called Deconverters. It is important to remark that these Deconverters were not developed in a uniform way. Some of them were done using the component provided by the UNL Center called "Deco" and some others reusing pre-existing Machine translation systems. This was the case of the Russian ETAP system and the French ARIANE. Other tasks of this first year were the construction of the UNL dictionaries based one the called Universal Words (UW) and the analyzers to convert the original language into UNL (1st UNL Symposium, UNESCO Headquarters Paris, November 1997).
- ?? **1998:** This was a year focused in the developing of the Second version of the UNL specifications, the generation and dictionary systems of all participant teams. This second year was for the developing of mature tools able to demonstrate the capabilities of the UNL system. Some UNL editors were showed at prototype level for some languages. First steps towards the creation of the UNL Society and the Technical and Quality Committees A public real demonstration was made during the 2nd UNL Symposium, UN Headquarters, New York, November 1998.
- ?? **1999:** Presentation of the UNL system to the press and to the translation services of the European Union. Specifications of the UNL language were open to the public by the requirement of the UN to avoid

exclusive uses of the UNL. The Quality Plan for the UNL organization was also produced. 3rd UNL Symposium, Brussels November 1999.

- ?? **2000-2001:** Advanced prototypes of encoding tools were developed by several teams. Definition of the complete architecture of the UNL System. Decision of the United Nations University for the creation of an independent organizations to take care of the developing, promotion and expansion of real applications for the UNL system. This new organization was the UNDL Foundation placed in Geneva, were the UNL Center should progressively be placed. First UNL-based applications were also developed and showed. Public presentations in international forums and developing of some real applications for public use. 4th UNL Symposium, UN Headquarters, Geneva, January 2001. Constitution of the UNDL Foundation. Fifth UNL open Symposium. Suzhou (China). November 2001.

2. Current State

After four years of sustained development, the UNL system comprises a set technical specifications, tools and resources:

- ?? Mature specifications of the UNL language and the Universal words (lexical entries). The definition of these elements is perhaps the greatest achievement of the project, since the specification of a language capable of accommodating widely different natural languages is considered an extremely difficult task. The specifications are in the public domain and available at www.undl.org and www.unl.ias.unu.edu
- ?? Mature generation systems for most of the languages initially handled by the project. During year 2000, several teams put their generation systems publicly available in Internet, building Language Servers that provide generation services.
- ?? Prototypes of several natural-language-to-UNL tools have been developed following different approaches, ranging from purely manual editors to fully automatic encoders.
- ?? UNL-adapted dictionaries have been compiled for all the languages involved.

Equally important, organizational developments have secured an institutional framework that guarantees the continuity of the project:

- ?? The UNDL Foundation plays the role of central co-ordinator and institutional representation of the UNL Programme on a world-wide scale.
- ?? The UNDL Foundation has approved the creation of twelve Language Centers, each one in charge of develop and promote the UNL within the boundaries of their respective language area.
- ?? The Technical and Quality Committees have been set up. Their mission is to promote a consistent use of the interlingua and to implant quality procedures.

3. Scope of the demonstration

The Demonstration has been organized in two parts. The first one is addressed to show how is the process to create the UNL representation from an original language. The second part will be around the use of the different tools already developed for the creation of UNL code and also to generate target languages from the UNL code produced. This process will be explained together the presentation of a public UNL demonstrator placed in Internet. A graphical introduction will be made in order to show the workflow of the complete process. We will explain in a detailed way these parts

3.1. First Part: Showing the process

The basic process for the creation of UNL from an original language will be explained, based on a graphical presentation (similar to the contents of a poster). The architecture, and the main components of the UNL language (Universal words, and relations) will be showed and explained. Printed materials will be distributed to the attendance.

In the fig. 1 we may visualize the basic architecture of the UNL System. The Analyzer (A) and the Deconverter or Generator (G) are integrated parts of this architecture. The UNL document base (UNL D) is also represented. The UNL written documents are stored in this component. The UNL Dictionary (DIC) is the lexical resource of the UNL where the called Universal Words are introduced. Following this representation the UNL basic relations, their role and functions will be described. T is the original language text and GT the target language Text (or Generated Text).

As any other semantic language, the UNL is described by the composition of a set of UWs, their attributes and the relations among the concepts. The basic approach to define a UW will be explained, creating some UWs and describing the structure of the UNL dictionary and its relations with the local dictionaries.

UNL is a partially formalized, artificial language for expressing informational content in a language-independent way. Essential components of this language are:

?? Universal Words (UWs): They represent concepts valid in all languages. Universal Words are composed by:

1. Basic UW: It is an English "headword" (for practical reasons) (e.g.: snow)
2. Restrictions: They perform a more accurate semantic definition of the UW. {e.g.: cat (icl>animal) or snow(icl>thing) vs. snow(icl>occur)}

?? Attributes: They modify the meaning of the Uws. They are:

- number (singular or plural); (e.g. @pl)
- reference (definite or indefinite); (e.g. @def)
- time and aspect;
- epistemic and deontic modalities;
- pragmatic intentions

?? Relations: Express the semantic relation between two UWs. There are:

- Semantic relations between UWs
- 41 conceptual relations, grouped into 8 families (according to the specifications)

Some examples of graphical representations will be showed and also the corresponding UNL code. After these explanations we will give the direction where lot of materials to support the manual learning of the UNL encoding, as <http://www.unl.fi.upm.es/CLE/> searching the called interactive pages where people may find materials, resources and real testing by the connection to the language servers of French, Russian, Italian and Spanish languages.

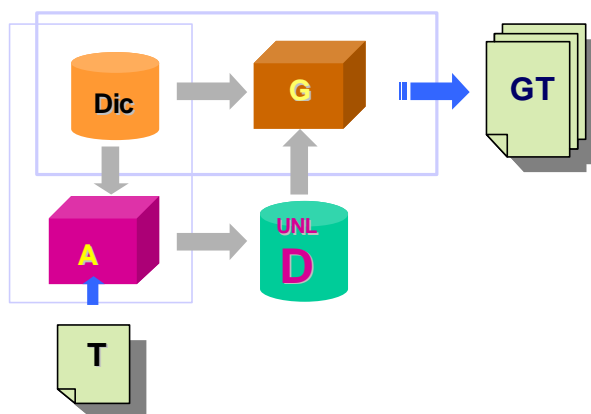


Figure 1. Basic architecture of the UNL System

3.2. Second part: use of the tools

The idea is to give a comprehensive picture of the whole process of creating a UNL document from a given text and generating the corresponding counterparts in several target languages. For the first process, we will demonstrate a UNL Editor developed for encoding Spanish texts into UNL. This tool follows a semi-automated approach, offering to its user a series of aids that speed up and validate the encoding process. The generation process will be demonstrated afterwards, showing the raw, non-edited generation results that come

out from several Language Servers. The encoding will be made from Spanish and English language into UNL starting from the same text. For that, we will use the UNL Editor developed by the Spanish Language Centre after a previous description of the functioning of the UNL Editor.

Additionally we will show how the UWs are introduced in the System and how they are critic to assure that the target language is well generated. Once produced the UNL code corresponding to the selected texts we will introduce it in the online language servers of each language (no less than four languages) in order to visualize the process.

Finally, we will show the called repository where several documents written in UNL produce the target language final documents as a simple example of the main characteristic of the UNL documents defining clear differences with other approaches. They basically are:

- ?? UNL is not only a language for computers, it is the language in which documents are written and stored. This means that those that have UNL written documents have the key to visualize them in other languages once the corresponding language deconverter has been developed. In this sense we could say that the UNL is a kind of knowledge repository.
- ?? The updating of the UNL documents is very simple. Needed changes suppose work only at level of the UNL code. The new generation is made only one time and stored.
- ?? A new business model appears: The exploiting of UNL written contents. New deconverters will access immediately to all the existing UNL documents.

We will complement the demonstration with a brief outline of the most promising applications of the UNL technology in areas such as E-commerce, multi-lingual web support and others, stressing the differences with current machine translation systems.

4. Acknowledgements

We acknowledge the co-operation in the preparation of this demonstration of the following research institutions and groups, all of them participants of the UNL Programme:

- ?? GETA-CLIPS, Université Joseph Fourier, CNRS, France
- ?? STAR SPB Ltd. Saint Petersburg, Russia
- ?? Institute for the Transmission Problems. Russian Academy of Sciences. Moscow
- ?? Istituto di linguistica Computazionale del CNR. Pisa (Italy)
- ?? UNL-Center, UNL Foundation, Geneva, Switzerland