# USING WORDNET FOR BUILDING AN INTERLINGUAL DICTIONARY[1]

Juan Bekios, Igor Boguslavsky, Jesús Cardeñosa, Carolina Gallardo

*Abstract: UNL is an enterprise to support multilinguality in Internet based on a common language called Universal Networking Language. One of the components of the language is a dictionary of Universal Words (UWs). Such dictionary constitutes the link between the vocabularies of the languages involved in the project. This article describes the process of creating the common UWs dictionary within the UNL context, using as an external resource Wordnet. The process is completely automatic. Implementation details and results of the process are shown.*

*Keywords: Lexical Resources, Wordnet.*

*ACM Classification Keywords: J.5. Arts and Humanities; H.2.8 Database Applications;*

## Introduction

The UNL Program was initiated by the Institute of Advanced Studies of the United Nations University under the UN auspices with an ambitious goal: to break down or at least to drastically lower the language barrier for the Internet users. It was launched in November 1996; the project embraced 14 groups from different countries representing a wide range of languages: Arabic, Chinese, German, French, Japanese, Hindi, Indonesian, Italian, Mongolian, Portuguese, Russian, Spanish and Thai.

The UNL Program pivots on the *Universal Networking Language*, a meaning representation language designed to represent informational content conveyed by natural languages. The complete specifications of the language are public and freely downloadable from Internet (see [Uchida et al., 2005]). One of the major applications of UNL is to serve as an interlingua between different natural languages. Besides that, UNL can also be used for other applications such as information retrieval, text summarization and the like. In fact, the specifications have known several versions, from version 1.0 in 1997 to current version of 2005, due to the fact that the language accommodates itself to new uses.

The UNL is composed of three main elements: universal words (UWs hereafter), relations and attributes. UWs form the vocabulary of the interlingua; relations express thematic roles and attributes represent the context and speaker dependent information. Formally, a UNL expression can be viewed as a semantic net, whose nodes are UWs, linked by arcs labeled with UNL relations. Universal Words are expanded by the attributes.

The complete set of UWs composes the UNL dictionary. The UNL dictionary is complemented with bilingual dictionaries, connecting UWs with words of different natural languages. Local dictionaries are formed by pairs of the form <Word, UW> where Word is any word of a given natural language and UW is the corresponding representation of one of its senses in UNL. The UNL dictionary constitutes a common lexical resource for all natural languages currently represented in the project, so that word senses of different natural languages become linked via their common UWs. Therefore, the UNL Dictionary can serve as an important lexical resource to

[1] This paper has been sponsored by the Ministry of Education and Science of Spain under project **CICYT HUM2005-07260**.

construct multilingual dictionaries or other resources like thesauri, being UWs the pivot among the vocabulary of natural languages.

However, there is an apparent drawback in the UNL dictionary. The set of UWs is not formally defined, that is, the specifications do not provide either a complete knowledge base or precise instructions to create UWs. The absence of formalization of the lexical part of the language prevents the construction of a common dictionary of UWs for all the members of the project, the management of dictionaries and lexical resources being the hardest part of the project.

This paper presents a methodology and an application that tries to solve the main problems in the UNL dictionary management, namely, the standardization of the definition of UWs and the automatic construction of the common UNL dictionary on the basis of the existing lexical resources.

## Data Analysis

As already said, UWs constitute the vocabulary of the language. Broadly speaking, a UW is an English word modified by a series of semantic restrictions. The main purpose of semantic restrictions is to eliminate lexical ambiguity present in natural languages. Besides that, they establish major lexical relations with other words and specify an argument frame. 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 argument frames. In a sentence like "*The plane landed at the Geneva airport*", the corresponding UW for the sense of this verb would be land(icl>do, plt>surface, agt>thing, plc>thing). This UW is divided in two parts: the headword and the list of semantic restrictions enclosed in parenthesis and separated by commas, as shown in figure 1.

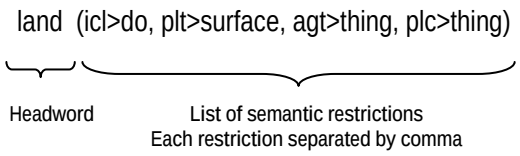


Figure 1. Parts of a UW

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 *plc* stands for *place*) establish the obligatory semantic arguments of the predicate "land".

As can be seen from this example, UNL semantic restrictions are based on lexical relations among terms, namely, the *hyponymy* relation (by means of "icl" relation), *synonymy* ("equ" relation) and *meronymy* ("pof" relation). Besides, the semantic arguments of predicates (that is, verbs, adjective and some nouns) must be specified. Since UWs are described by means of relations between terms, the result is a connected net of UWs, constituting the UW system. A more comprehensive view of the UW system is described in [Boguslavsky et al, 2005].

The organizing principles of the UW system are based on well-known lexical relations, like those present in Wordnet [Fellbaum, 1998]. Wordnet is a large lexical database of English, freely downloadable from Internet

(http://wordnet.princeton.edu/). As opposed to most lexicographic works and similarly to the UW system, Wordnet is not ordered alphabetically but conceptually, by means of semantic relations. The main organizing relation in Wordnet is the synset, defined as a group of cognitive synonyms that expresses a single concept. Besides, synsets are interconnected by means of other lexico-semantic relations like hyperonymy (hierarchical relation between class and subclass), antonymy (an opposite term), metonymy (part-of) and other relations like *relative_to*, sentence frames for verbs, etc. Figure 2 shows two samples of Wordnet that illustrate the relations of hyperonymy and antonymy for the synset "male child, boy".

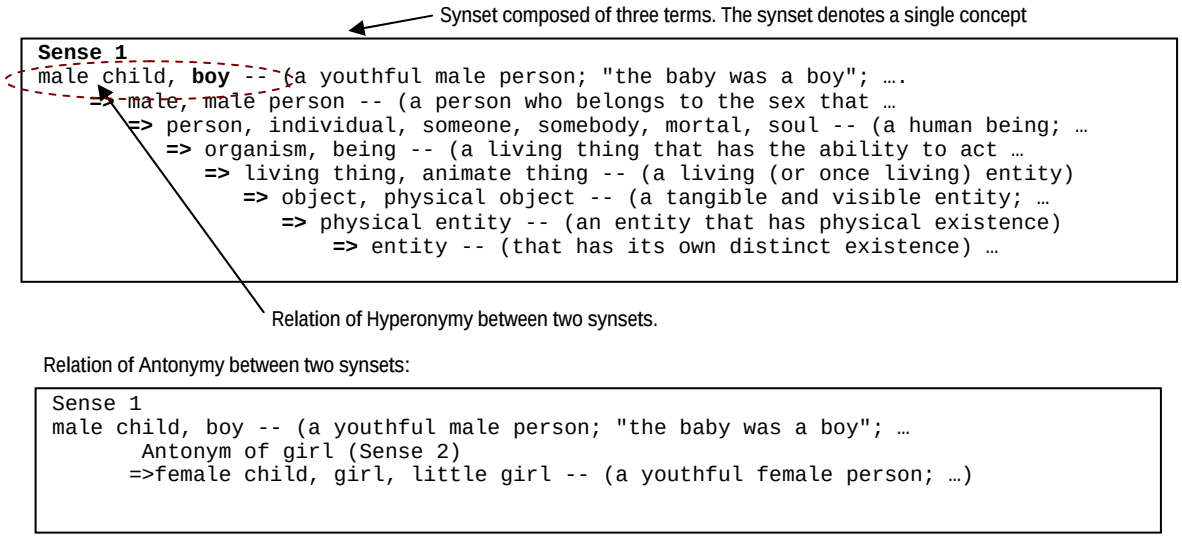


Fig 2. Two samples of Wordnet

Wordnet includes nouns, adjectives, adverbs and verbs. Other categories like prepositions, determiners or conjunctions are spelled out from Wordnet, since they do not denote any semantic concept.

The use Wordnet as an ancillary resource to support the process of automatic dictionaries creation is not new in the UNL framework. The generation of UNL-English dictionaries for specific texts is depicted in [Bhattacharyya et al, 2004]. We have made use of the similarity of Wordnet and the UW system to use Wordnet as the main source to define and create a complete UW dictionary. The complete process and the final UW dictionary are described in the next sections.

## Design Issues

The main design issue when considering a UW Dictionary and Wordnet as the main source of data is that the structure of lexical relations in Wordnet can be used to construct the list of restrictions of UWs. To do that, we must first establish the main similarities between Wordnet and the UW system. Such similarities are exposed in table 1, where the first column describes lexical relations in Wordnet and the second column states their equivalent semantic restrictions in the UW system.

Table 1 shows how any word included in Wordnet can be used to represent the headword of a UW. Each different sense of an English word is delimited by means the set of synonyms, hypernyms, antonyms and other lexical relations associated to that word, in the same way that the sense of a headword in UNL is delimited by its list of semantic restrictions.

| SIMILARITY RELATIONS | |
|---|---|
| WordNet 2.1 | UW System |
| An English Word. | Headword |
| Synset | Relation equ> |
| Hyperonym | Relation icl> |
| Antonym | Relation ant> |
| Relative to | Relation com> |

Table 1. Similarity Relations between Wordnet and the UW system

What is really important for us is that from these similarity relations, it is possible to devise a method that defines UWs in a systematic way using Wordnet. The method is described in figure 3.

```
1. Extract a Word from Wordnet
2. Obtain each of the senses of the Word
3. For each sense of the word, do the following:
    3.1. Assign the Word to the Headword of UW
    3.2. Depending on the syntactic category (noun, adjective, adverb, verb)
         and on the data obtained from WordNet; for each sense, apply a set
         of rules that will generate semantic restrictions.
    3.3. Taking the Headword and the obtained restrictions, construct the
         complete Universal Word.
    3.4. Store the UW in the dictionary.
4. If more UWs are to be constructed, return to step 1. Otherwise, finish.
```

Fig. 3. Method to define UWs from Wordnet

There are two aspects that require further explanations in this method. First, the number of UWs that are created per word and second the set of rules mentioned in step 3.2 of the method.

The method will generate one UW per word sense. For example, the word “bank” as a noun has 10 senses and thus generates 10 different UWs. In some cases, when the difference between the senses is too subtle, Wordnet relations are not sufficient to differentiate between them. In these cases, the method will generate identical UWs for different senses. These “duplicate” UWs must be treated in a special way.

On the other hand, the method is based on the similarity relations of table 1 along with a set of rules to systematically yield a dictionary of UWs. These rules are presented in the next section.

## Set of Rules

Only six rules are required to create the semantic restrictions of UWs. A rule takes as input a Wordnet word (that is, the set of senses for the word and the lexical relations each word is engaged in) and yields a semantic restriction suitable for the UW that is being created. The six rules are:

1. Rule for the Construction of Headword (HW)

   *Definition:* This rule turns a WordNet word into a Headword for a candidate Universal Word.

   *Example:* The word “*banking company*” in Wordnet returns the Headword “***banking_company***”.

2. Immediate Hypernym Rule (RHper)

   *Definition:* For a sense of a word, take its most immediate hypernym and establish an icl> relation type.

   *Example:* For the first sense of the word “*bank*” as a noun, take its immediate hypernym (“*financial institution)* and create a semantic restriction with icl>. The result is: “icl> financial_institution”

3. Immediate Hyponym rule (RHpo)

*Definition:* For a sense of a word, take its most immediate hyponym and establish an icl< relation type. Use this relation only when there are duplicate UWs.

*Example:* For the first sense of the word "*bank*" as a noun, it is possible to obtain navigating through WordNet an immediate hyponym ("for example *credit_union")* and create a semantic restriction with icl<. The result is: "icl<credit_union"

4. Rule of First Synonym (RSyn)

*Definition:* For a sense of a word, if the word is not the first element of the synset, take the first word of the synset and establish an equ> relation.

*Example:* For the first sense of the word "*bank*", it synset is {*depository financial institution*, *bank*, *banking concern*, *banking company*}. Since "bank" is not the first element, create the following semantic restriction: "equ> depository_financial_institution"

5. Rule of First Antonym (RAnt)

*Definition*: For a sense of a word, take its associated antonym (if any) and establish an "ant>" relation.

*Example:* For the adjective "*good*" in its first sense, the antonym associated to its first sense is "*bad*", therefore the generated restriction will be: "ant>bad"

6. Rule of Relative_to (RRel)

*Definition:* For a sense of a word (usually adjectives), take the associated noun by means of relation "pertains to" (if any) and establish an "com>" relation

*Example:* For "the *legal*" adjective, WordNet establishes a relation belongs to the noun "*law*", therefore the following restriction is obtained: "com>law"

These rules are independent of each other and can be executed in any order. When constructing the complete UWs dictionary, the application of rules will depend on the syntactic category of the headword (that is, not all rules are relevant for a given syntactic category). For example, when working with verbs, the application of the Antonym rule is irrelevant, since the meaning of a verb is not characterized by its antonyms. Table 2 summarizes the rules that are triggered for each syntactic category.

| *Syntactic category* | *Executed rules* |
|---|---|
| Noun | HW, RHper, RSyn, RAnt |
| Adjective | HW, RHper, RAnt, RSyn, RRel, |
| Adverb | HW, RSyn, RAnt, RRel |

Table 2. Set of rules relevant for each syntactic category

That is, a given noun may produce at most 4 semantic restrictions. For example, the noun "boy" in its first sense produces the following semantic restrictions:

- icl>male>thing (by means of RHper)
- equ>male_child (by means of RS)
- ant>girl (by means of RA)

The final UW is the concatenation of the generated semantic restrictions following the same order of table 2:

boy(icl>male>thing, equ>male_child, ant>girl)

The order of semantic restrictions implicit in table 2 is a convention followed by all the team members of the project. A different ordering will not imply different semantics of the UW.

Verbs are treated in a different way. Whereas all the information required for creating good UWs for nouns, adverbs and adjectives is present in the Wordnet, the mapping between verbal UWs and verbs in Wordnet is not so straightforward. This is due to the following reasons:

- Verbal UWs are categorized into three basic types of events: “do”, “occur” and “be”. This categorization is absent in Wordnet.
- Verbal UWs should be provided with its semantic arguments. Verbs in Wordnet are assigned a Sentence Frame, which is a, often incomplete, description of syntactic arguments for verbs.

Since there is no one-to-one relation between verbal UWs and the verbs, it was necessary to infer the type of event and the semantic arguments from the scarce information present in Wordnet. For that, we made use of the so-called lexicographic files which define broad ontological categories. Some of these categories are “verbs of dressing and bodily care”, “cognition verbs”, “verbs of being and having”. The combination of the ontological category together with the sentence frame of a verb gives us a hint about its type of event and semantic arguments. Table 3 shows an excerpt of the combinations that have been used to define verbal UWs.

| Wordnet | | UNL | | |
|---|---|---|---|---|
| Ontological category | Sentence Frame | Event Type | Semantic Arguments | Example |
| verbs of being, having, spatial relations | Somebody ----s to somebody | be | aoj>thing,obj>thing | conform(icl>be,aoj>thing,obj>thing) |
| verbs of weather | Somebody ----s | occur | obj>thing | steam(icl>occur,obj>thing) |
| verbs of creation | Somebody ----s something | do | agt>thing,obj>thing | cut(icl>do,agt>thing,obj>thing) |

Table 3. Combinations to define verbal UWs.

## The Dictionary Application

The complete application is composed of the following modules, graphically shown in figure 4:

- Conversion Module: This component converts words from Wordnet into UWs. This module uses the Rules and the Wordnet data. The generated Universal Words are served to the Database Manager.
- Database Manager: This component manages all the communications to and from the Database. Thus, this module receives the set of generated UWs from the Conversion Module and serves them to the Database. On the other hand, this module manages the processes of searching, modifying, deleting and inserting UWs as requested by users through the Web Browser. This component was developed in Java, using the special library Hibernate. (www.hibernate.org). In the near future, the UW Dictionary is expected to store the translations of UWs not only into English but into the other languages of the project.
- Web Browser: It refers to any existent web browser like Explorer, Firefox, Opera, etc. which will be used by users in order to interact with the UW Dictionary.

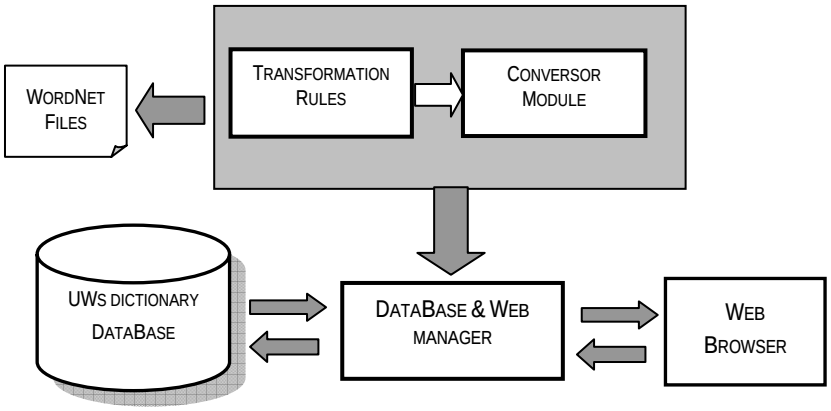


Fig 4. Components and relations of dependency of the Dictionary of Universal Words

The application can be accessed at the following address: http://chueca.dia.fi.upm.es:8080/UNLDicWeb/

## Results

All the UWs of the resulting UW Dictionary have been created automatically, without human intervention. Obtained results for a total amount of 207016 words that have been processed are summarized in table 4, where the total amount of generated UWs divided in syntactic categories is shown. The percentage of duplicate UWs for each syntactic category is also specified.

| | Nouns | Adjectives | Adverbs | Verbs |
|---|---|---|---|---|
| Unique UWs | 142343 | 26784 | 4958 | 23716 |
| Duplicate UWs | 2761 | 4518 | 762 | 1174 |
| Total | 145104 | 31382 | 5728 | 24890 |
| % duplicate UWs | 1,9% | 14,39% | 13% | 4,7% |

Table 4. Obtained results

Since nouns are by far the most elaborated category in Wordnet, we considered as correct UWs the set of unique UWs, and as incorrect the set of duplicate UWs. As can be seen from table 4, the rate of duplicate UWs for nouns is less than 2%, a good result for the most polysemous syntactic category. Surprisingly, the results for verbs is rather good (less that 5% of error rate), although we assume that semantic arguments of verbs require human revision. On the other hand, both adjectives and adverbs yield an error rate quite high (around 14%). The possible reason for such an error rate may lie in the fact that the main lexical relations present in Wordnet are synonymy and hypernym, natural relations for nouns but not for predicates like adjectives or adverbs.

## Bibliography

[Bhattacharyya et al, 2004]. N. Verma and P. Bhattacharyya, *Automatic Lexicon Generation through WordNet,* Global WordNet Conference (GWC-2004), Czech Republic. Jan, 2004

[Boguslavsky et al, 2005]. Boguslavsky, I., Cardeñosa J., Gallardo, C., and Iraola, L. The UNL Initiative: An Overview. Lecture Notes in Computer Science. Volume 3406/2005, pp 377-387. Springer Berlin / Heidelberg: 2005. ISBN 978-3-540-24523-0

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

[Uchida et al, 2005] Universal Networking Language (UNL). Specifications Version 2005. Edition 2006. 30 August 2006. http://www.undl.org/unlsys/unl/unl2005-e2006/

## Authors' Information

Igor Boguslavsky – Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: igor@opera.dia.fi.upm.es http://www.vai.dia.fi.upm.es

Juan Bekios – Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: juan.bekios@opera.dia.fi.upm.es. http://www.vai.dia.fi.upm.es

Jesús Cardeñosa – Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: carde@opera.dia.fi.upm.es. http://www.vai.dia.fi.upm.es

Carolina Gallardo – Group of Validation and Industrial Applications. Escuela Universitaria de Informática. Universidad Politécnica de Madrid. Carretera de Valencia Km.7. 28041 Madrid; email: cgallardo@eui.upm.es. http://www.vai.dia.fi.upm.es