The conceptual structure of the Encyclopedia of Water: remarks from a UNL enconverting task

Ronaldo Martins

UNDL Foundation
Geneva, Switzerland
r.martins@undlfoundation.org

ABSTRACT
This paper analyzes the results of translating, from English into the Universal Networking Language (UNL), 25 articles of the Encyclopedia of Water, one of the several encyclopedias of the EOLSS. We present the statistics for UWs, relations and attributes, which are the building blocks of UNL, and address some general issues about the use of UNL for knowledge representation and extraction.

Keywords
UNL, knowledge representation, information structure.

1. INTRODUCTION
The Universal Networking Language (UNL) is an "electronic language for computers to express and exchange every kind of information" [1]. It can be defined as a knowledge representation technique expected to figure either as a pivot language in multilingual machine translation systems or as a representation scheme in information retrieval applications. It has been developed since 1996, first by the Institute of Advanced Studies of the United Nations University, in Tokyo, Japan, and more recently by the UNDL Foundation, in Geneva, Switzerland.

In the UNL approach, information conveyed by natural language is represented, sentence by sentence, as a hypergraph composed of a set of directed binary labeled links (referred to as "relations") between nodes or hypernodes (the "Universal Words", or simply "UW"), which stand for concepts. UWs can also be annotated with "attributes" representing information that cannot figure as UWs or relations.

Formally, a UNL statement is a semantic network believed to be logically precise, humanly readable and computationally tractable. For instance, the English sentence 'Peter kissed Mary?!' can be represented in UNL as follows:

```
[S]
{unl}
agt(kiss(agt>person,obj>person).@entry.@past.@interrogative.@exclamative,
Peter(iof>person))
obj(kiss(agt>person,obj>person).@entry.@past.@interrogative.@exclamative,
Mary(iof>person))
{/unl}
[/S]
```

In the example above, 'agt' (agent) and 'obj' (object) are relations; 'Peter(iof>person)', 'Mary(iof>person)' and 'kiss(agt>person,obj>person)' are UWs; and '@entry', '@past', '@interrogative' and '@exclamative' are attributes.

Differently from other semantic networks, such as conceptual graphs [2] [3], and the RDF [4], UNL is not only a formalism; it is an entire language, enclosing a lexicon (the set of UWs) and a grammar (the set of relations and attributes). As of the version 3.3 of the UNL Specifications [5], the set of binary relations, which is supposed to be closed and permanent, consists of 46 semantic cases (such as agent, object, instrument, etc); the set of attributes consists of 72 elements (interrogative, imperative, polite, etc); and the set of UWs, which is open and subject to increase, consists of more than 60,000 entries.

In the UNL framework, the process of representing information into UNL is called "enconversion", and the process of extracting natural language sentences out of UNL is called "deconversion". For the time being, the enconversion process can be defined mostly as a computer-assisted human analysis of natural language into UNL, while the deconversion has been rather a fully-automatic generation from UNL into natural language, with some human post-editing.

2. EOLSS
EOLSS (an acronym for Encyclopedia of Life Support Systems) is the one of world's largest online publications dedicated to the Natural and the Social Sciences. Available at http://www.eolss.net, it is an integrated compendium of more than twenty encyclopedias, which attempts "to forge pathways between disciplines and to foster the transdisciplinary relations between subjects especially related to the life supporting systems" [6].

As a product of thousands of experts from over 100 countries, EOLSS has been facing some shortcomings related to its knowledge management structure:

1) it is monolingual: all articles have been published only in English;
2) it is unidimensional: articles are not hypertexts (i.e., they do not contain hyperlinks to other texts, except for the section "related chapters"); and
3) it is poorly standardized: the metadata, for instance, is not uniform, and the same authors ("Karl Steininger" and "Karl W. Steininger") or institutions ("Tokyo University" and "University of Tokyo") appear differently in different articles.

In order to improve the access and to normalize EOLSS, the UNDL Foundation has proposed to represent it as a knowledge network to be formalized in UNL. The first stage of this process, which has been setting the guidelines for the overall enterprise, concerns the UNLization of 25 articles selected randomly from the Encyclopedia of Water, one of the many encyclopedias of EOLSS. These 25 articles comprise 12,917 sentences, 199,983 tokens (words) and 12,878 words (types) of English, which have been already enconverted into UNL by the UNL Center, located in Tokyo, Japan. The results of the enconversion are available at the project website (http://www.undlfoundation.org/eolss) and constitute the main inspiration for this paper.

3. RESULTS

The results of the enconversion of the EOLSS corpus are summarized in Table 1.

Table 1. Results of the enconversion of 25 articles of EOLSS

UNL	
Isolated Nodes	1,303
Relations (tokens)	118,731
Relations (types)	42
Attributes (tokens)	164,448
Attributes (types)	39
UWs (tokens)	238,255
UWs (types)	15,532

The enconversion found 1,303 isolated nodes (i.e., single-node graphs), which are one-word titles and subtitles, and correspond to 10% of the enconverted sentences. In general, there was an average of 10,22 relations per sentence, comparable to the average of 15,53 words per sentence in the original corpus.

The distribution of the use of relations is shown in Figure 1.

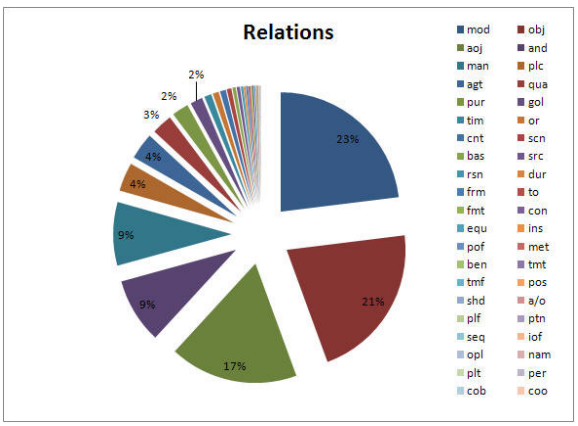

Figure 1. Distribution of relations in the EOLSS corpus

The "mod" relation (23%), which is mainly used for nominal modifiers, such as "water supply" = mod(supply, water); the "obj" (21%), which is used for complements, such as "to create a tsunami" = obj(create, tsunami), or "after the agreement" = obj(after, agreement), as well as for the subject of event verbs, such as "the temperature decreases" = obj(decrease, temperature); and the "aoj" relation (17%), which is used for noun predicates, such as "Arab world" = aoj(Arab, world), and for the subject of state verbs, such as "the extract contains" = aoj(contain, extract), are responsible for more than 60% of the occurrences. In total, 26 relations were used more than 100 times, and 42 relations were used at least one time. Interestingly, 4 relations ("cag", for co-agents; "cao", for co-attributes; "icl", for hyponymy; and "int", for interjection) were not used. Whereas the latter two are actually expected to figure only in ontologies (along with "iof" = instance of , "equ" = equal to, and "pof" = part of, which did appear in the corpus), the first two are supposed to describe ordinary thematic roles and it was quite unexpected that they were not encountered in such a comprehensive corpus.

The distribution of the use of attributes brought many discrepancies as well. They are illustrated in Figure 2. The attribute ".@entry", which has no semantic value but is mandatory for every UNL sentence and subsentence (scope), was the most frequent one. The second one was the attribute for plural ("@pl), with 25% of the occurrences, followed by the attribute "@def" (17%), which normally replaces the definite article ("the"). The attribute "@topic", which is used mainly to indicate a syntactic inversion (as in passive constructions) was the fourth (12%), followed by the attributes "@past", for the past tense, and "@indef", for the indefinite article, both with 5%.

The proportions, however, can be misleading, in the sense that the attribute "@progress", which corresponds mainly to the present continuous of English verbs, and which was the 7th most used, had actually 5,287 occurrences. As a matter of fact, 21 attributes (listed below from "@entry" to "@interrogative") occurred more than 100 times. Five others (from "@ordinal" to "@title") were used more than 10 times, but almost 40 attributes present in the UNL Specs were not used at all.

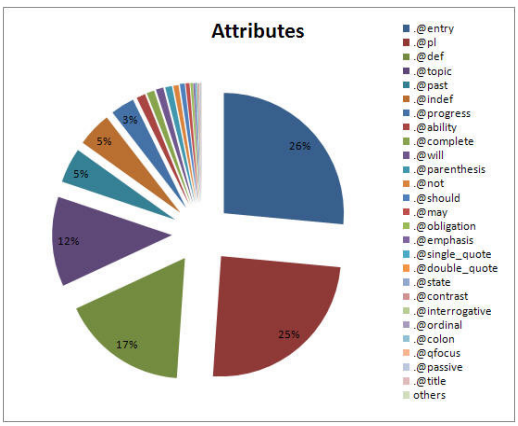

Figure 2. Distribution of attributes in the EOLSS corpus

As for the UWs, only 4 were used more than 1,000 times: "water(icl>liquid)" (8,949 times), "use(agt>thing,obj>thing)" (1,508 times), "use(icl>act)" (1,081 times) and "it(icl>thing)" (1,039 times). Other 357 UWs appeared more than 100 times, and 3,453 appeared only once. The whole set of UWs, with their frequency of occurrence, can be found at the project website, referred to above. As depicted in Figure 3, most UWs are nouns (54%), followed by adjectives (20%) and verbs (18%).

.

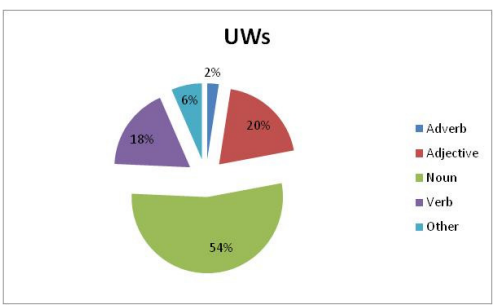

Figure 3. Distribution of UWs in the EOLSS corpus

4. THE CONCEPTUAL STRUCTURE OF THE ENCYCLOPEDIA OF WATER

We believe that the raw data presented in the previous section points to the conceptual structure of the Encyclopedia of Water. The analysis of the 10 most frequent words, presented in Table 2, for instance, clearly indicates that the subject of the corpus has to do with the "use", the "quality" and the "resources" of "water", as well as with the "methods" and the

"processes" for water "supply", particularly in relation to the "soil". This is exactly the subject of the Encyclopedia of Water.

Table 2. The 10 most frequent UWs in the EOLSS corpus

UW	Frequency
water(icl>liquid)	8949
use(icl>act)	1081
quality(icl>attribute)	878
resource(icl>functional thing)	811
method(icl>way)	717
soil(icl>substance)	661
process(icl>event)	646
area(icl>place)	605
supply(icl>act)	605
system(icl>structure)	587

The same results, however, could be easily obtained by simply counting the nouns in the original text. The advantages of having the corpus in UNL start to appear when we consider the use of relations and attributes. No direct processing of the original corpus would lead us to notice, for instance, that:

a) the role of "agents" in the corpus have been remarkably omitted, what can be derived from the high frequency of the attribute "@topic" and the low frequency of the relation "agt";
b) besides being primarily assertive (what can be perceived from the low frequency of the attributes "@interrogative", "@not" and others indicating modality), the corpus is mostly descriptive (what is indicated by the high frequency of "aoj" and the significant presence of ontological relations, such as "equ", "iof", "pof" and "cnt"); and
c) in addition to complex sentences (10,22 relations per sentence), the corpus is mainly comprised by a qualified vocabulary (what is indicated by the high frequency of the relation "mod", combined with the high frequency of the attribute "@def").

Those stylistic considerations tell a lot about the nature of the corpus and the rhetorical structure of its texts, and can definitely be used as a strategy for text proofing and subject indexing. Nevertheless, the most outstanding use that can be made from such data concerns perhaps information networking. The whole collection of documents has been represented as several different semantic networks (one per sentence) which can be interlinked by their common nodes in order to constitute a single complex system that is very suitable for knowledge engineering.

Let us consider, for example, the result of interlinking 10 different UNL sentences bringing the UW "water(icl>liquid)", which has already been referred as the most frequent one. The results for the most frequent occurrences of "water" as the target node of an "aoj" relation whose source node is an adjective are presented in Table 3 below.

Table 3. The 10 most frequent "aoj" relations between an adjective and "water(icl>liquid)" in the EOLSS corpus

RELATION	FREQUENCY
aoj(potable(aoj>thing), water(icl>liquid))	96
aoj(virtual(aoj>thing), water(icl>liquid))	45
aoj(fresh(aoj>water), water(icl>liquid))	43
aoj(green(aoj>thing), water(icl>liquid))	36
aoj(ultra-pure(aoj>thing), water(icl>liquid))	33
aoj(blue(aoj>thing), water(icl>liquid))	26
aoj(available(aoj>thing), water(icl>liquid))	26
aoj(pure(aoj>thing), water(icl>liquid))	19
aoj(industrial(aoj>thing), water(icl>liquid))	17
aoj(clean(aoj>thing), water(icl>liquid))	11

From the Table 3 it is possible to build a whole map structure of the properties of "water": it can be "potable", "virtual", "fresh", "green", "ultra-pure", "blue", "available", "pure", "industrial" and "clean". Additionally, it can also be said, at least according to the corpus, that "water" it is likely to be "green" rather than "blue", "fresh" rather than "clean", and so on. The potentialities of such networking range from extracting "semantic collocations" that are not normally registered in ontologies to extracting complex definitions of terms, especially if we consider that several other subsidiary relations can be retrieved in the same corpus. The relation between "green" and "blue", for instance, which is quite important for the definition of "water", can be inferred from the 18 occurrences of the relation "and(green(aoj>thing), blue(aoj>thing))" that appear in the corpus.

5. FINAL REMARKS

Despite of the results achieved so far, it is interesting to observe that the enconversion of the 25 articles of the Encyclopedia of Water also poses several issues to the UNL Specifications themselves, which seem to require a further revision in order to enhance its strengths.

The distribution depicted in Graph 1 shows clearly that there are relations (such as "mod", "obj" and "aoj") that have been underspecified, in the sense they are currently covering different types of semantic phenomena, whereas there are others (such as "cag" and "cao") that may have been overspecified, as they have not been used at all. The same applies to attributes, which have been used rather scarcely, given that 50% of the possiiblities listed in the UNL Specifications have not been used one single time in a corpus that is reasonably comprehensive.

The UWs involve different problems: the analysis of the corpus proves, once again, the need for better standards, as there seems to be a gratuitous proliferation of labels for the same concepts ("blue(aoj>color)" and "blue(aoj>thing)", for instance), as well as an excessive dependency of the English vocabulary (the UW "it(icl>thing)", which is one of the most frequent in the corpus, is a noteworthy example: it actually represents an index rather than a concept, and should have been represented in a different way).

In any case, the results substantiate the idea that UNL can be used for several different purposes, including translation, and encourages the improvement of the current technology of the UNL system. Indeed, the EOLSS corpus has been used, inside the UNDL Foundation, as the main current case study, and has oriented an extensive reconsideration of both the enconversion and the deconversion processes, with the development of new tools. The main idea is that representing a document in UNL can be not only a strategy for multilingualization, but also for knowledge reverse engineering, which can have many interesting applications, such as multi-document summarization, concept mining and semantic proofing.

REFERENCES

[1] Uchida, H., Zhu, M. and Della Senta, T. (1999) A gift for a millennium, IAS/UNU, Tokyo.
[2] Sowa, J. F. (1984), Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, Reading, MA.
[3] Sowa, J. F. (2000), Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA.

[4] Lassila, O. and Swick, R. R. (1999) Resource Description Framework (RDF): model and syntax specification. W3C Recommendation.
[5] UNL Centre. (2005). UNL Specification. Version 3.3. UNL Centre/UNDL Foundation, Geneva.
[6] Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO, Eolss Publishers, Oxford ,UK, [http://www.eolss.net].