

Knowledge Vertices in XUNL

Ronaldo Martins

Abstract—This paper addresses some lexical issues in the development of XUNL – a knowledge representation language descendent from and alternative to the Universal Networking Language (UNL). We present the current structure and the role of Universal Words (UW) in UNL and claim that the syntax and the semantics of UWs demand a thorough revision in order to accomplish the requirements of language, culture and human independency. We draw some guidelines for XUNL and argue that its vertices should be represented by Arabic numerals; should be equivalent to sets of synonyms; should consist of generative lexical roots; should correspond to the elementary particles of meaning; and should not bear any non-relational meaning.

Index Terms—Knowledge representation, UNL, lexical resources, semantic networks.

I. INTRODUCTION

THE XUNL (eXtended UNL) is a knowledge representation language descendent from, and alternative to, the Universal Networking Language (UNL). It was inspired by the UNLX Project [1] – an intensive experience of using UNL for natural language analysis and generation – and by the resulting conviction that some principles of UNL should be thoroughly reformulated in order to cope with multilingual knowledge representation.

In short, XUNL departs from four axioms:

- I. Knowledge can be represented as the directed hyper-graph $K = (V, H)$, where V is the set of knowledge vertices, and H – which stands for knowledge hyper-edges – is a set of non-empty subsets of V .
- II. A hyper-graph $K = (V, H)$ is isomorphic ($\simeq$) to the hyper-graph $K' = (V', H')$ if there exists a bijection $\phi: V \rightarrow V'$ and a permutation π $\phi(h_i) = f_{\pi(i)}$, where $h_i \subseteq H$.
- III. A natural language utterance describes a sub-hyper-graph of K induced by a subset A of V , such as $KA \subseteq K$ and $KA = (A, (H \cap A)^*)$.
- IV. Two natural language utterances KA and KB are interchangeable if and only if they are isomorphic ($K_A \simeq K_B$).

In this paper, we address the first premise, which states that human knowledge is discrete and relational: it would consist of “knowledge vertices” linked by “knowledge hyper-edges”. Knowledge vertices (KV) would stand for atomic facts of human thought; and knowledge hyper-edges (KH) would

represent directed relations between knowledge vertices or, recursively, between other knowledge edges.

In order to define the nature and the role of KVs, which roughly correspond, in the UNL approach, to the Universal Words (UWs), we revise the structure and the function of UWs in UNL, and demonstrate that the syntax and the semantics of the KVs should be rather different, in order to accomplish the requirements of language, culture and human independency. We draw some lexical guidelines for XUNL and prescribe the structure of KVs.

The paper is organized as follows: Section 2 presents briefly the Universal Networking Language; Section 3 brings the current syntax and semantics of UWs; Section 4 analyzes the concept of Master Definition (MD) and the structure of the UNL Knowledge Base (the UNLKB); Section 5 explores some problems and shortcomings of the current version of the UNL KB; and finally Section 6 introduces the guidelines for KVs in XUNL out of our experience with UNL.

II. UNL

The Universal Networking Language (UNL) is an “electronic language for computers to express and exchange every kind of information” [2]. 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 at the United Nations University, in Tokyo, Japan, and more recently by the UNDL Foundation, in Geneva, Switzerland.

Formally, UNL is a semantic network believed to be logically precise, humanly readable and computationally tractable. In the UNL approach, information conveyed by natural language is represented, sentence by sentence, as a hyper-graph composed of a set of directed binary labeled links (referred to as “relations”) between nodes or hyper-nodes (the “Universal Words”, or simply “UW”), which stand for concepts. UWs can also be annotated with “attributes” representing mainly modality.

As a matter of example, 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]
```

Manuscript received on October 31, 2011, manuscript accepted on December 9, 2011.

Ronaldo Martins is with UNDL Foundation, 48 Route de Chancy, Petit-Lancy, CH-1312 Geneva, Switzerland (e-mail:r.martins@undlfoundation.org).

Differently from other semantic networks (such as conceptual graphs [3, 4] and the RDF [5]), 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 2005 UNL Specifications [6], the set of relations, which is supposed to be closed and permanent, consists of 44 binary relations (such as agent, object, instrument, and so on); 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 63,000 entries.

Under the UNL Program, natural language analysis and understanding is referred to as a process of “enconverting” from natural language (NL) into UNL. This enconverting process, which has been carried out in a computer-aided human basis, is said to be not only a mere encoding, but truly a translation from the source sentence into a new target language – UNL – which is claimed to be as comprehensive as any NL. As a matter of fact, and at least for the time being, UNL has been mainly used for multilingual document generation, through a process referred to as “deconverting”, which consists in automatically providing NL outputs to hand-coded UNL graphs.

III. UW

Universal Words, the words of UNL, are composed of a root (usually referred to, in UNL Specifications, as “headword”) and a suffix (“the constraint list”). The latter comes between parentheses and is used mainly to disambiguate the former. Examples of UWs are presented below:

- (1a) ‘Universal Word’
- (1b) ‘UW(equ>Universal Word)’
- (1c) ‘Peter(iof>person)’
- (1d) ‘apple(icl>fruit)’
- (1e) ‘kiss(agt>person,obj>person)’
- (1f) ‘explain(icl>express(agt>thing,gol>person,obj>thing))’
- (1g) ‘Manyoshu(icl>Japanese poem)’

In order to be mnemonic and humanly readable, roots and suffixes are labeled out of English words, except for culture-dependent concepts (1g above). The UNL Center does not take that to be a language bias and claims that UWs are only arbitrary labels: they are but unique strings of characters used to refer to concepts. The meaning of a UW would be entirely derived from the so-called UNL Knowledge Base (or simply UNLKB), a huge network where nodes are interconnected as to emulate the structure of human cognition.

As a matter of example, the meaning of “apple(icl>fruit)” should be defined by a set of binary relations such as those indicated by (2) below :

- (2a) *icl(apple(icl>fruit),fruit(pof>plant))=1;*
- (2b) *obj(eat(agt>thing,obj>thing), apple(icl>fruit))=1;*
- (2c) *aoj(round(aoj>thing), apple(icl>fruit))=1;*
- (2d) *pof(apple(icl>fruit), apple tree(icl>tree))=1;*

Actually, in the UNL Program, there seems to be at least two different representational levels for defining UWs. The

first is related to the UNLKB itself and targets the (alleged) systematic part of the meaning, in a sense very close to the one intended by the concept of “semantic markers” [7]. On the other hand, the unsystematic part of meaning (the “distinguishers”) is treated in the UNL Encyclopedia, which is a huge UNL document base, also organized as a network, where idiosyncrasies and additional information on UWs are expected to be stored. Here we will focus only on the UNLKB structure.

IV. UNL KNOWLEDGE BASE

The UNLKB is a semantic network in which entries have the structure exemplified in (2) above. They comprise a binary directed relation (extracted from the UNL relation set) between two UWs, along with a degree of certainty, which can range from 0 (completely false) to 255 (completely true). Any UNL relation can hold between UWs in the UNLKB, and a single UW may receive and assign many different relations from and to other UWs. However, in order to guarantee inference and cross-reference inside the network, every UW should be linked to another one by at least one of three ontological relations: “icl” (a-kind-of), “iof” (an-instance-of) or “equ” (equal-to).

Linking one UW to another by means of “icl”, “iof” or “equ” is to compose a sort of thesaurus, the UW Ontology, which is part of the UNLKB. Inside the UNL System, this subnetwork has been referred to as the “UW System”, and constitutes a lattice structure, given that a single child-node may have many different parent-nodes. This hierarchical network also comprises an inheritance mechanism, so that all information assigned to a given parent-node can be directly inherited by its children-nodes. In this sense, if (3) below had been stated in the UNLKB, there would be no need for (2b), provided that it could be easily inferred from (2a):

- (3) *obj(eat(agt>thing,obj>thing), fruit(pof>plant))=1;*
- (2a) *icl(apple(icl>fruit),fruit(pof>plant))=1;*
- (2b) *obj(eat(agt>thing,obj>thing), apple(icl>fruit))=1;*

The need for the UNLKB has been subject to criticism inside the UNL Project, but it should be observed that knowledge-based MT systems have proved to provide better results than those that are only language-based [8]. Inside the UNL System, the UNLKB is intended to assure robustness and precision both to the NL-UNL enconverting and to the UNL-NL deconverting. In the former case, the UNLKB would be used as a sort of word sense disambiguation device; in the latter, the UNLKB, through replacement operations, would allow for the deconversion of UWs not enclosed in the target language dictionaries. Additionally, the power of the UNLKB for intelligent search and semantic reasoning should never be underestimated.

In order to discipline and organize the creation of UWs, the UNL Center has proposed a particular technique for both naming and defining a UW in a single movement: the Master Definition (MD), introduced in 2000. The MD for naming the UW “apple(icl>fruit)” and defining it in the UNLKB (through an “icl” relation to the UW “fruit(pof>plant)”) is presented in (4) below:

(4) *apple(icl>fruit{pof>plant})*

The MD is said to facilitate (and regulate) the labeling of a UW, which would derive its suffix (the constraint list) from its definition in the UNLKB. The name of the UW would simply be the same as the MD without the strings inside the curly braces.

However, it should be noticed that the concept of MD brings itself at least two serious shortcomings: 1) due to the simplification of syntax, the MD is not capable of conveying any degree of certainty other than 1; and 2) MDs can only be used to define the UW by means of ‘icl’, ‘equ’ or ‘iof’; any richer definition would require longer strings and more expensive strategies. Nevertheless, and at least for the time being, the UNLKB has been entirely defined as a hierarchy of MDs.

V. PROBLEMS

We shall here concentrate on three main problems concerning the set of UWs in the current status of the UNL Program. The first is language-dependency; the second is culture-dependency; the third is human-dependency. In all those cases, we will address the set of UWs available at www.undl.org as of July, 2011. Some changes have been provided since then, but the problems remain basically the same.

A. Language-dependency

As for language-dependency, we claim that lexicalization of UWs has been exaggeratedly based on lexical items of English. This can be attested by the extensive presence of English idiosyncrasies in the set of UWs.

For instance, one will find, in the UNLKB, both “behavior(icl>action)” and “behaviour(icl>action)”. The difference between them is not semantic, but strictly orthographic, and there is no reason for cataloging such kind of spelling difference in a semantic database.

The same should apply for pairs of antonyms such as give/receive, borrow/lend, etc. These verbs are supposed to convey the same meaning in a reversed subcategorization frame: give(x,y) = receive(y,x). Once “give” and “borrow” are there, would there be any reason for including “receive” and “lend” as well?

(5a) *give(agt>thing,gol>person,obj>thing)*

(5b) *receive(agt>thing,obj>thing,src>thing)*

(6a) *borrow(agt>thing,obj>thing)*

(6b) *lend(agt>thing,gol>person,obj>thing)*

This sort of overlapping among UWs does not affect only antonyms and can be found all over the UNLKB. Let us consider two last examples: is there any real need for registering, in the same knowledge base, all the words appearing in (7) and (8) below? Are the semantic differences between them really relevant? Are they going to be preserved in languages other than English?

(7a) *begin(agt>thing,obj>thing)*

(7b) *commence(icl>begin(agt>thing,obj>thing))*

(7c) *start(icl>begin(agt>thing,obj>thing))*

(8a) *nurse(icl>medical assistant)*

(8b) *nurse({icl>person>human,}icl>occupation{>work})*

The examples referred to above prove that economy has not been an asset of the UNLKB. Obviously, one may claim that variation should be represented, because there is no perfect synonymy, and UNL is supposed to be as comprehensive and fine-grained as any natural language. However, in this case, we would have a problem even more severe: provided that there is no perfect lexical matching between languages, UNL would have to register every word from every language, what would not only degrade the performance and the maintenance of UNL resources, but lead UNLKB to entropy and solipsism.

B. Culture-dependency

Culture-dependency can be detected mainly in the UNLKB categorization procedures, which has involved many inconsistencies. Tigers and panthers, for instance, are normally defined as belonging to the species of felines, but, in the UNLKB, they have been categorized directly under “mammal(icl>animal)”, differently from “cat(icl>feline)”:

(9a) *tiger(icl>mammal{>animal})*

(9b) *panther(icl>mammal{>animal})*

(9c) *cat(icl>feline{>mammal})*

In the same way, specific languages and types of languages have been categorized at the same level, as indicated in (10) below:

(10a) *spoken language{(icl>language>system)}*

(10b) *Russian(icl>language{>system})*

(10c) *inflectional language{(icl>language>system)}*

Circularity may also be found, as in (11) and (12):

(11) *thing{(icl>nominal concept)}*

abstract thing{(icl>thing)}

event(icl>abstract thing{>thing})

thing(icl>event{>abstract thing})

(12) *figure(icl>figure{>attribute})*

The main problem concerns the lack of criteria for categorization. In (13) below, for instance, the concept conveyed by the English words “film” and “movie” is said to be linked to the concept of “abstract thing”. Why that? Why not “concrete thing”? Or why not “functional thing”? What about instances of films, such as “Gone with the wind”? Would they also be considered a kind of “abstract thing”?

(13) *abstract thing{(icl>thing)}*

art(icl>abstract thing)

cinema(icl>art{>abstract thing})

film(icl>cinema{>art})

movie(icl>cinema{>art})

Such categorization turns out to be even more astonishing if we consider the case for “book”, which is also located under the “abstract thing” branch of the UNLKB, as indicated in (14):

(14) *abstract thing{(icl>thing)}*

information{(icl>abstract thing)}

document(icl>information)

TABLE I.
ENGLISH-TO-PORTUGUESE CORRESPONDENCE FOR THE NOUN “BOOK”

English	Definition	Portuguese
1. book	a written work or composition that has been published printed on pages bound together; "I am reading a good book on economics"	livro
2. book, volume	physical objects consisting of a number of pages bound together; "he used a large book as a doorstep"	brochura
3. ledger, leger, account book, book of account, book	a record in which commercial accounts are recorded; "they got a subpoena to examine our books"	registro
4. book	a number of sheets ticket or stamps etc. bound together on one edge; "he bought a book of stamps"	álbum
5. record, record book, book	a compilation of the known facts regarding something or someone; "Al Smith used to say, 'Let's look at the record'"; "his name is in all the recordbooks"	registro
6. book	a major division of a long written composition; "the book of Isaiah"	livro
7. script, book, playscript	a written version of a play or other dramatic composition; used in preparing for a performance	livro
8. book, rule book	a collection of rules or prescribed standards on the basis of which decisions are made; "they run things by the book around here"	livro
9. Koran, Quran, al-Qur'an, Book	the sacred writings of Islam revealed by God to the prophet Muhammad during his life at Mecca and Medina	Livro
10. Bible, Christian Bible, Book, Good Book, Holy Scripture, Holy Writ, Scripture, Word of God, Word	the sacred writings of the Christian religions; "he went to carry the Word to the heathen"	Livro

book(icl>document{>information})
book of general works{(icl>book>document)}
manuscript{(icl>book of general works)}
rare book{(icl>book of general works)}
book of geography{(icl>book>document)}

On the other hand, both “landscape” and “scenery”, and even “beauty spot”, are categorized under “concrete thing”, as seen in (15):

(15) *concrete thing{(icl>thing,icl>place>thing)}*
natural world{(icl>concrete thing,icl>place>thing)}
landscape(icl>natural world)
scenery(icl>landscape{>natural world})
beauty spot(icl>scenery{>landscape})
scene(icl>scenery{>landscape})

The absence of categorization guidelines causes the UNLKB to be excessively impressionist, in the sense it contains, to a considerable extent, subjective and personal ideas towards the world and the structure of events. Although some of those decisions may sound quite reasonable from a given perspective, it is clear that they cannot be taken for granted. They are rather culture- and even individual-dependent and will be subject to an everlasting dispute. In fact, this is said to be the main reason why knowledge-based approaches have been discarded as a feasible strategy for language processing and, inside the UNL Program, this is probably the reason why there is so much resistance on adopting a more fine-grained level of lexical description.

Actually, outside the UNL Center, it has been observed a relatively flat use of the suffixes of UWs, as if their only role was to assign some part-of-speech information to the roots. As a result of that, UWs such as “book(icl>thing)” have been

more frequent than “book(icl>document)”, for instance. These simple UWs, however, are not trouble-free either: they are not able to totally disambiguate English words and to assure precision and robustness to both enconverting and deconverting. In the Princeton WordNet [9], for instance, the noun ‘book’ (presented in Table I), may take 10 different senses, some of which may not be translated, in Portuguese or in any other language, by the same single word. In those circumstances, a low-level use of suffixes would not only be insufficient, but mostly misleading. To reduce all senses of “book” to “book(icl>thing)” would be no better than declaring that “book” is a sort of “abstract thing”.

Consequently, the best solution for the limitations pointed out above is not to deprive the UNLKB, restricting its power and the granularity of its representation. Actually, the answer is to keep improving the UNLKB, but in a rather different perspective, as suggested in the next section.

C. Culture-dependency

The third limitation to be addressed here concerns the alleged “human-readability” of UNL graphs. As indicated above, UNL is an “electronic language for computers to express and exchange every kind of information”. It is not a language for humans. The argument that UNL graphs should use English words because they would be human-friendly is not only pointless but contradictory. Only very specialized people would be capable of reading UWs, and no one would be actually able to understand them, provided that UWs are only labels whose meanings should be extracted from the UNLKB. As we have already stressed, “apple(icl>fruit)” does not stand for “apple” in the human sense; it is the set of all relations

departing from and coming to “apple(icl>fruit)” in the UNLKB.

The option for English words, therefore, is not only useless but mainly misleading and deceiving. One should never forget that, from the computer point-of-view, “apple(icl>fruit)” is nothing but a memory address which will be meaningful if, and only if, is associated to other several memory addresses. If UNL is to be taken not as a mere notation, not as a mere natural language script, but as a real language, a truly different and autonomous one, self-coherent and natural-like, which could figure either as a source or a target in machine translation systems, we should strengthen that UNL cannot be about other languages (such as English), but that UNL must be directly about what other languages are about.

VI. SOLUTIONS

In order to circumvent the problems depicted in the last section, we propose three radical modifications in the UNL approach: UNL should not imitate English; UNL should not be an ontology-based language; and UNL should not be a knowledge-based language. As they strongly deviate from the main beliefs of the UNL approach, we acknowledge that such changes will lead us to something that is no longer UNL. In order to avoid confusion and dispute, we have been using the name “eXtended UNL” (or simply “XUNL”). XUNL keeps the core idea that knowledge conveyed by natural language sentences could be represented by hyper-graphs, in which nodes would stand for concepts, and edges would consist of directed binary semantic relations. However, nodes and edges in UNL and in XUNL are remarkably different, and should be therefore differently referred to. XUNL nodes have been addressed as “Knowledge Vertices” (KV) instead of UWs; and XUNL relations have been called “Knowledge Hyper-edges” (KH). In this section, we trace some general guidelines for Knowledge Vertices.

A. Language-dependency

The first commitment – not to imitate English – can be understood in two different senses. The most easily achievable is that XUNL should no longer use English words, or that KVs should be made out of language-independent symbols, such as Arabic numerals. In this case, KVs would not be as readily legible as UWs, but would be shorter, less deceptive and actually universal. Additionally, human readability could be easily provided by editing facilities as these existing in the very computer where this text is being typed, which automatically converts Roman characters into machine-tractable codes. Indeed, there is no actual need for middle-level representations (such as UWs and MDs) in the current state of the art of human-machine interfaces.

However, the language-independency commitment must also be understood in a far much deeper and much more intricate way. It is not only a matter of labeling, but of choosing what is supposed to be labeled. Spelling differences (‘color’ and ‘colour’) and synonyms (‘freedom’ and ‘liberty’) should clearly not be represented as different lexical items in XUNL. The set of KVs should be equivalent to the set of synonyms of a given language instead of to the whole set of

words of that language. In this sense, KVs would be very akin to the concept of synset devised by the WordNet [10].

Moreover, XUNL should comprise only lexical roots (monomorphemic stems), i.e., the set of atomic lexical items necessary and sufficient to generate the whole set of words of a given language. For instance, there is no need, in XUNL, for a word like “beautiful” or “beautifully”, provided that we have “beauty” and some derivation rules. This is to say that the XUNL lexicon should be generative, instead of enumerative.

Finally, XUNL should include only the semantic elementary particles of lexical meaning. Natural language words should be represented as complex semantic structures to be analyzed in XUNL. Accordingly, a verb like “to fly” should be rather represented as “to travel through air”, (or even more radically as “to change location through air”), and a noun like “chair” should be represented as “a seat for one person, with a support for the back”. Natural language lexical items should not be simply translated in XUNL but truly defined in relation to a core minimum vocabulary, as simple and small as possible.

B. Culture-independency

There seems to be compelling evidence that human knowledge is not organized in tree-like deep hierarchies but in a quite different topology, consisting of prototypes, exemplars and family resemblance relationships that are mainly contingent (context-dependent) rather than essential or necessary [11, 12]. It is ineffective for instance, to state that “apple” is a kind of “fruit” because, in many different contexts, it is not. In sentences like (16) below, where only a part of the meaning of apple is actually activated, knowledge that apples can be considered types of fruits is not only superfluous but even misleading.

(16) *This ball of yam looks like an apple.*

The fact is that concepts are exceedingly complex and fluctuating structures that, due to analogical reasoning, often assume unpredictable meanings. It is hopeless, therefore, and often useless, to build static ontologies, which will never be able to portrait the markedly different structure of human knowledge. Instead, the set of KVs should constitute a huge distributed network of associations molded by experience rather than logic.

In order to avoid categorization biases and ontology shortcomings, the set of KVs should be defined, not in a knowledge base, but in an example base (a “memory”) automatically extracted out of real texts. This memory could bear, in principle, the same formal structure of the UNLKB, i.e., a set of directed binary relations between KVs associated to a given degree of “certainty”, which would have to be understood as “frequency of occurrence”; however, it would be formed not out of human insights on classes and classification principles, but out on actual co-occurrence in a given corpus. Additionally, this memory would have to be incremental, since the frequency of occurrence would be permanently revised as new data are processed.

The two major issues in the XUNL Example Base are the corpus and the machine learning strategies. As for the former, we shall admit that the larger the better. The ideal corpus would comprise every available document, so that relations

between KVs would be set as broadly as possible. However, it is unlikely that one would be able to process so many data in order to extract recurrent edges between KVs. A more realistic approach would recommend the notion of “archive”, in a sense very close to the one intended by Foucault [13], or that of “norm”, according to Coseriu [14]. In both cases, we would acknowledge the fact that some texts are more authoritative than others, and should have, therefore, a more prominent role in validation.

The second problem – related to (unsupervised) machine learning – has already received considerable attention in Artificial Intelligence and can be easily adapted to the process of mechanically extracting co-occurrence relations in a given corpus. Clustering [15] and neural networks [16] have proved to exemplify interesting possibilities for pattern extraction and classification in large amount of data.

C. Human-independency

At last, we should stress that XUNL will be able to embody human knowledge and to occupy the places of source and target language in machine translation systems if, and only if, XUNL is self-consistent and human-independent. In order to “mean” to a machine, XUNL should not mean to a human being, who operates in a completely different way, from a completely different structure.

XUNL should not rely on definitions derived from human comprehension or an external language that is not replicable by the machine, but, instead, should shape its own world, a purely intensional (non-mental) dimension, a sort of electronic (possible) world, which would represent the sense and the reference of XUNL words and expressions. This digital (and artificial) world, and not the human analogical one, would figure as the “aboutness” of XUNL, and would comprise the truth-condition requirements for XUNL expressions to be “meaningful”.

This is to say that the set of KVs should constitute a sort of sign system where the value of a given sign should solely derive from its position in the network. At least at the lexical level, XUNL should consist of “un système où tout se tient” [17], following hence the structuralist approach that “every language is a system, all parts of which organically cohere and interact [... where] no component can be absent or even different, without transforming the whole” [18].

Accordingly, KVs would only bear a negative (relational) definition, which would not necessarily coincide with the positive normally ascribed by a human. The value of a given KV would be the sum, and nothing but the sum, of all relations in which it takes part in the XUNL Example Base.

VII. FINAL REMARKS

It may seem clear by now that this paper is rather prescriptive than descriptive. Our main goal here was not to describe the existing syntax and semantics of Knowledge Vertices, but to draw some general guidelines that should be used in building them. The rules are seven and they are the following:

- I. *KVs should be represented by Arabic numerals (instead of English words).*
- II. *KVs should be equivalent to sets of synonyms (instead of individual words).*

- III. *KVs should be equivalent to generative lexical roots (instead of inflected or derived forms).*
- IV. *KVs should be equivalent to the elementary particles of meaning (instead of complex semantic structures).*
- V. *KVs should not be organized (or defined) according to single human-crafted ontology.*
- VI. *KVs should be organized (and defined) according to unsupervised machine learning procedures operating over a selective example base.*
- VII. *KVs should not bear any positive (non-relational) meaning.*
- VIII. *Implementing such a lexical database is the first step and the first challenge in the development of XUNL.*

REFERENCES

- [1] www.ronaldomartins.pro.br/unlx
- [2] H. Uchida, M. Zhu, and T. Della Senta, *A gift for a millennium*, IAS/UNU, Tokyo, 1999.
- [3] J. F. Sowa, “Conceptual Structures: Information,” in *Processing in Mind and Machine*, Addison-Wesley, Reading, MA, 1984.
- [4] J. F. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA, 2000.
- [5] O. Lassila and R.R. Swick, *Resource Description Framework (RDF): model and syntax specification*, W3C Recommendation, 1999.
- [6] UNL, www.unl.org/unlsys/unl/unl2005/
- [7] J. Katz and J. Fodor, “The structure of a semantic theory”, *Language*, 39, pp. 170-210, 1963.
- [8] S. Nirenburg, V. Raskin et al., “On knowledge-based machine translation,” in *Proceedings of the 11th International Conference on Computational Linguistics*, Bonn, 1986.
- [9] *WordNet*, wordnet.princeton.edu
- [10] C. Fellbaum, *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press, 1998.
- [11] D. L. Medin and A. Ortony, “Psychological Essentialism,” in S. Vosniadou, A. Ortony, (eds.) *Similarity and Analogical Reasoning*, Cambridge: Cambridge University Press, 1989.
- [12] E. Rosch, “On the Internal Structure of Perceptual and Semantic Categories”, in T. E. Moore (Ed.), *Cognitive Development and Acquisition of Language*, New York: Academic Press, 1973.
- [13] M. Foucault, *L'Ordre du discours*, Paris, Gallimard, 1971.
- [14] E. Coseriu, *Teoría del lenguaje y lingüística general*, Madrid, Gredos, 1962.
- [15] S. Kotsiantis and P. Pintelas, “Recent Advances in Clustering: A Brief Survey,” *WSEAS Transactions on Information Science and Applications*, Vol 1, No 1 (73-81), 2004.
- [16] G. Hinton and T. Sejnowski (eds.), *Unsupervised Learning and Map Formation: Foundations of Neural Computation*, MIT Press, 1999.
- [17] F. de Saussure, *Cours de Linguistique Générale.*, Paris, Payot, 4 ed. Trad. bras. São Paulo, Cultrix, 1969.
- [18] H. G. Gabelenz, *Die Sprachwissenschaft*, 1901.