Universal Words and their relationship to Multilinguality, Wordnet and Multiwords

Pushpak Bhattacharyya[1]

Abstract:
In this article we address issues concerning construction of lexicon in the context of sentential knowledge representation in *Universal Networking Language (UNL)*, an interlingua proposed in 1996 for machine translation. Lexical knowledge in UNL is in the form of *Universal Words (UWs)* which are concepts expressed by mostly English words disambiguated and stored in the universal words repository. The UW dictionary is universal in the sense that it aims to store all concepts of all languages of all times. After many incarnations of UW dictionaries were built at many places of the world, the UW dictionary that is based on the content and structure of English wordnet seems to have found unanimous acceptance in the UNL community. We point that some of the concerns that challenge the wordnet building activity, challenge the construction of the UW dictionary too. These concerns are multilinguality and multiwords. We discuss possible solutions to these challenges, noting on the way that challenges to UW dictionary construction arise from the ambition of UWs repository to be universal.

1. Introduction

Natural Language Processing (NLP) has assumed great importance in today's world due to the proliferation of the Web. Huge quantities of text in electronic form are available on the internet, waiting to be processed and made sense of. As a task, NLP is three dimensional, involving *languages*, *problems* and *algorithms*, as shown in figure 1 (Bhattacharyya, 2012).

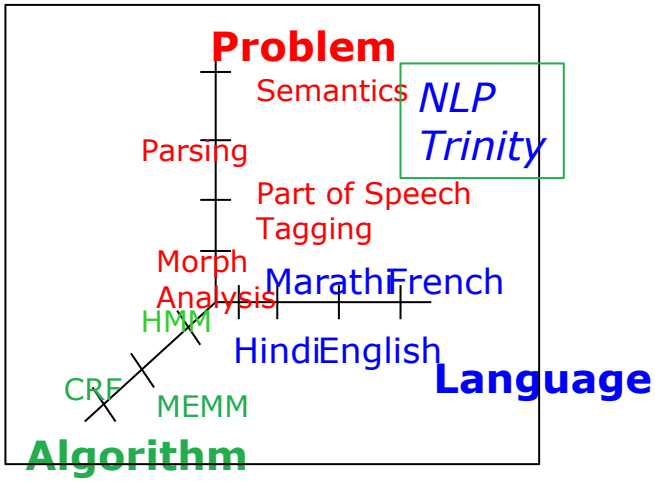

Figure 1: NLP Trinity

For example, in Part of Speech Tagging for English using Hidden Markov Model (HMM), the language is English, the problem is POS tagging and the algorithm is HMM. NLP is also a

[1] Department of Computer Science and Engineering, IIT Bombay, India. pb@cse.iitb.ac.in

layered process (figure 2), since a set of tasks at level *l* need to be done before tasks at the next level *l+1* can be undertaken.

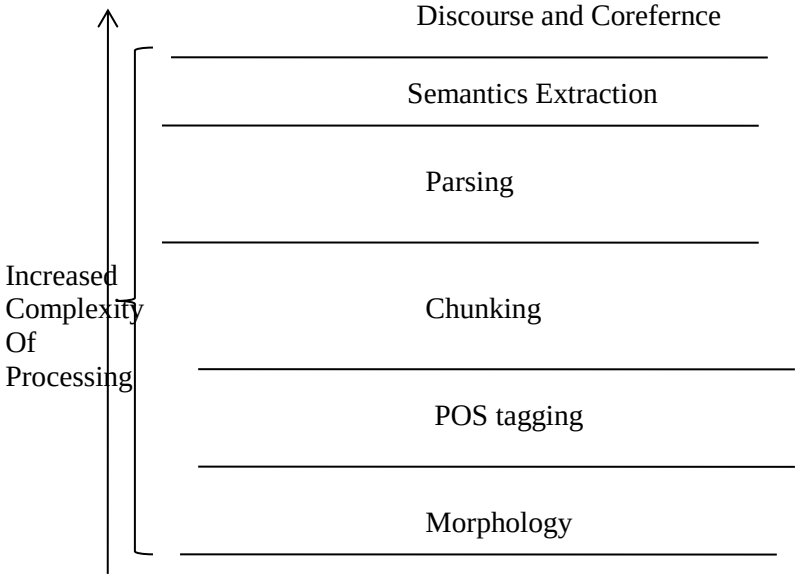

Figure 2: NLP layers

Of course, the processing of language cannot be strictly upward from one NLP layer to the next, because of *ambiguity*. Ambiguity processing is at the heart of Natural Language Processing. At every layer of NLP, choices have to be made as to what the label of a part of the input text (words, phrases, clauses etc.) should be. Multiple passes are required over the layers to finally arrive at the meaning of a text. This is similar to what happens in a complex compiler like GCC[2], where more than 200 passes finally produce the correct and optimized code.

Any NLP system has to have a clean, well defined framework for representing lexical knowledge, that is, *words* and their *meanings*. By words one does not mean only single character strings, but also hyphen and space separated multi-constituent text units, called *multiwords*.

The other reality of NLP in modern times is multilinguality, caused by the multilingual web. Taking the example of India, there are 4 major language families in the country, viz., *Indo European, Dravidian, Sino Tibetan* and *Austro-Asiatic*. Some languages are ranked within 20 in the world in terms of the populations speaking them: Hindi and Urdu: 5th (~500 milion), Bangla: 7th (~300 million), Marathi 14th (~70 million) and so on.

In what follows, in section 2, we first describe the existing system for representing lexical knowledge, called *Universal Words*, in Universal Networking Language[3]. Section 3 is on how to make the UW dictionary cater to demands of multilingality. Section 4 explores the

[2] http://gcc.gnu.org/

[3] UNDL Foundation. Universal networking language (unl) specifications, edition 2006, August 2006. http://www.undl.org/unlsys/unl/unl2005-e2006/.

relationship of multiwords with UWs. Section 5 compares and contrasts wordnet and UW dictionary. Section 6 concludes the paper, pointing to future directions.

2. Universal Words in Universal Networking Language

Universal Networking Language (UNL) is a knowledge representation scheme (Uchida et. al., 1999), very similar to *semantic nets (*Woods, 1975*).* Sentences are represented as graphs, wherein nodes represent concepts and directed edges represent semantic relations. The nodes are called *Universal Words* which express concepts unambiguously. *Relations* specify the roles of words in the sentence. *Attributes* stand for speech acts like *emphasis, time, plurality, aspect* and so on.

For illustration, consider the sentence "*Robots are used to find avalanche victims"*, the UNL graph for which is shown in figure 3. The concepts involved are *use, robot, find, victim* and *avalanche.* The relations are *obj (object), pur (purpose),* and *mod (modifier).* Here, "*find avalanche victim"* is an embedded concept which is represented as a hyper-node called *a scope*[4] node. The UNL graph of a sentence captures the lexical, syntactico-semantic and pragmatic content via UNL UWs, UNL relations and UNL attributes respectively. The predicates in the UNL graph of a sentence represent the atomic facts in the sentence.

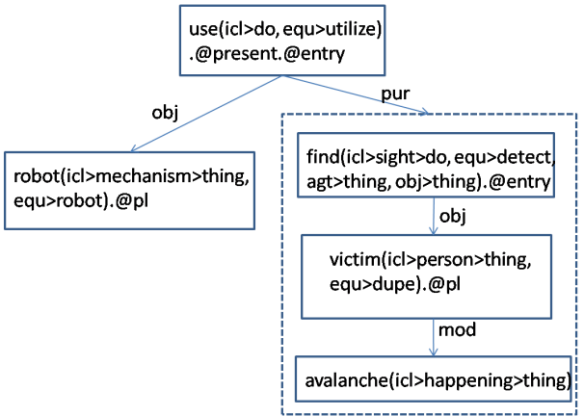

Figure 3: UNL Graph for the sentence *Robots are used to find avalanche victims* (taken from RTE-3 Development Set: Pair Id 270)

In the above figure, the main verb *use* has as its object *robots* (*obj* denotes object relation and *@pl* denotes plurality); the purpose (*pur* semantic relation) of using robots is a *finding* activity. The object of *find* is *victims*. The victims are of a special kind, viz., *avalanche victims* as indicated by the *mod* (modifier) semantic relation. *@entry* is a special attribute indicating the main predicate of the sentence/clause.

2.1 Universal Words (UWs) (Dave et al 2002)

UWs are made up of a character string (usually an English-language word) followed by a list of restrictions. When used in UNL expressions, a list of attributes and often an instance ID follow these UWs. The Head Word is an English word or a phrase or a sentence that is interpreted as a label for a set of concepts. This is also called a **basic UW** (which is without

[4] http://www.undl.org/unlsys/unl/unl2005/UW.htm

restrictions). For example, the basic UW *drink*, with no constraint list, denotes the concepts of "putting liquids in the mouth", "liquids that are put in the mouth", "liquids with alcohol", "absorb" and so on.

When a language→UW dictionary is constructed, for example Hindi→UW dictionary, language specific words written in the script of the language is linked with UW. For example, in

[पीना] "drink(icl>consume>do, agt>thing, obj>thing)"; take in liquids

the Hindi lexeme पीना (piinaa) written in Devanagari script is linked to the UW expressing the common concept of ingesting a liquid.

In BNF notation, the definition of an UW during its appearance in an UNL expression is:

<UW>::=<head word>[<constraint list>][: <UW ID>][. <attribute list>]

The constraint list restricts the interpretation of a UW to a specific concept. The restricted UW "drink(icl>consume>do, agt>person, obj>thing)" in the example above denotes the concept of "putting liquids into the mouth". "icl>consume>do" puts the concept in the category of "consume" and then in the category of "do", i.e., *transitive* verbs. Further disambiguation is done by invoking the argument frame cum selectional preference ("agt>person" and "obj>matter", i.e., "drink" requires an *agent* is of type *person* and an *object* is of type *matter*). Words from different languages are linked to these disambiguated UWs and are assigned syntactic and semantic attributes. This forms the core of the lexicon building activity in UNL.

An obvious question that arises for the UWs is "Why call these universal, since they are based on English?" As Katz says:

> Although the semantic markers are given in the orthography of a natural language, they cannot be identified with the words or expressions of the language used to provide them with suggestive labels. (Katz, 1966:156)

This means that the primitives exist independently of the words used to describe, locate or interpret them. The UWs, though represented using Roman characters and English lexemes, are actually language-independent concepts. For example, for the Hindi word □□□□ (*devar)* the English meaning is 'husband's younger brother'. We keep the universal word "husband's younger brother(icl>relative)" in the Hindi–UW dictionary and link it to *devar*.

It should be noted that the headwords in UWs are not always English words. Roman letters are used to represent all the concepts that are found in all the languages at all times. Thus, *ikebana* (a Japanese art form for folding papers) and *kuchipudi* (an Indian dance form) which are not English words are also stored in the dictionary.

Restrictions play the crucial role of sense disambiguation. They are written in Roman letters. But they do not depend on English. The senses are not the ones that are peculiar to the English language. For example, one of the senses found in India of the word "back bencher" is "student who is not serious in his/her studies and whiles away the time sitting at the back of the class". This additional sense is included in the UW dictionary as "back-bencher(icl>student)". Thus if a particular word *w* in English has acquired an additional sense in another language, this sense is introduced into the UW dictionary by tagging the appropriate restriction. The words in specific languages get mapped to specific word senses and not to the basic UWs. The basic UWs are ambiguous and the linking process is carried out only after disambiguating.

3. UWs and multilinguality (Dave et al 2002)

We have given the example of *devar* ("Husband's younger brother") in Hindi. This illustrates the case where there is no direct mapping from a language to an English word. We have to discuss the reverse case where for an English word there is no direct mapping in another language. This is important since the UWs are primarily constructed from English lexemes. The normal practice is that if an English word is commonly used in a language, say, Hindi, we keep the Hindi transliterated word in the dictionary. For example, for the word "mouse" used in the sense of an input device for the computer we store the following in the lexicon:

[माउस] "mouse(icl>device)"

The same strategy is adopted if a word is very specific to a language and culture. For example, for the English word "blunderbuss" (an old type of gun with a wide mouth that could fire many small bullets at short range), there is no simple Hindi equivalent and so we keep the transliteration in the lexicon.

[ब्लण्डरबस] "blunderbuss(icl>gun)";

The topic of multiple words for "snow" in Eskimo languages is very popular in the NLP, MT and Lexical Semantics literature. In the Eskimo language *Inuit*, the following are a few examples for the word "snow": "aput*"* means *snow (in general),* "pukak" means *snow (like salt),* "mauja" *soft deep snow*, "massak" *soft snow* and "mangokpok" *watery snow*. The rich set of relations of UNL is exploited to form the UWs which in this case respectively are shown as:

[aput] "snow(icl>thing)";
[pukak] "snow(aoj<salt like)";
[mauja] "snow(aoj<soft, aoj<deep)";
[massak] "snow(aoj<soft)";
[mangokpok] "snow(aoj<watery)";

Note the disambiguating constructs for expressing the UWs. The relations of UNL are used liberally. *aoj* is the label for the adjective–noun relation.

The issue of shades of meaning is a very important one, and again the relations of UNL can be used. Below we show are some of the shades of meaning of the verb *get off* and the noun *shadow* and the way of representing them. (The gloss sentences are attached for clarifying the meaning, which anyway gets communicated through the restrictions) .

[बचना] "get off(icl>be saved)"; lucky to get off with a scar only

[भेजना] "get off(icl>send)"; Get these parcels off by the first post

[बन्ध करना] "get off(icl>stop)"; get off the subject of alcoholism

[काम रोकना] "get off(icl>stop,obj>work)"; get off (work) early tomorrow.

For "shadow" which has many shades of meaning:

[अन्धेरा] "shadow(icl>darkness)"; the place was now in shadow

[काला धब्बा] "shadow(icl>patch)"; shadows under the eyes.

[परछाई] "shadow(icl>atmosphere)"; country in the shadow of war

[रंचमात्र] "shadow(icl>iota)"; not a shadow of doubt about his guilt

[साया] "shadow(icl>close company)"; the child was a shadow of her mother

[छाया] "shadow(icl>deterrant)"; a shadow over his happiness

[शरण] "shadow(icl>refuge)"; he felt secure in the shadow of his father

[आभास] "shadow(icl>semblance)"; shadow of power

[भूत] "shadow(icl>ghost)"; seeing shadows at night

Again, note should be made of how the restrictions disambiguate and address the meaning shade.

4. UWs and multiwords

Multiwords is a relatively new term (Sag et. al., 2002). A group of words that have a non-compositional meaning and/or have a fixity of lexeme and structure (collocation) are called multiwords. We regard the following to be the necessary and sufficient conditions for multiwordness:

A. A multiword must consist of space separated words (necessary condition)
B. A multiword should have (sufficient condition)
    a. Non-compositionality of meaning
    b. Fixity of expression
        i. In lexical items
        ii. In structure and order

For illustration of the necessary condition, consider the following Marathi sentence:

M1: Marathi: सरकार हक्काबक्का झाले
R1: Roman: sarakAra HakkAbakkA JZAle
E1: English meaning: *the government was nonplussed*

Here the string हक्काबक्का is a compound and not a multiword, since there is no space between the two components हक्का and बक्का. The following, however, is a multiword.

H2: Hindi: गरीब नवाज़
R2: Roman: garIba navAjZa
E2: English meaning: one who nourishes the poor

For the sufficient condition, following are examples of sufficiency arising out of non-compositionality:

K3: Konkani: पोटांत चाबता
R3: Roman: poTAMta cAbatA (literally, *biting in the stomach*)
E3: English meaning: to feel jealous

T4: Telugu: చెట్టు కిందికి పల్లీడరు
R4: Roman: ceVttu kiMda pLIdaru (literally, *a lawyer sitting under the tree)*
E4: Meaning: an idle person

B5: Bangla: □□□□□ □□□□□
R5: Roman: mAtira mAnuSa
E5; English meaning: a simple person/son of the soil

In all these examples, it is impossible to derive the meaning of the whole unit from the individual meanings of constituents.

For multiwords arising from fixity of lexical items, the examples are:

H6: Hindi: उसने मुझे गाली दी
R6: Roman: usane muJe gAlI dI
M6: English meaning: *he abused me*

But not,

H7: Hindi: * उसने मुझे गाली प्रदान की
R7: Roman: *usane muJe galI pradAna kI
M7: English meaning: *he abused me*
E8: **lifelong imprisonment* (*life imprisonment* is idiomatic)

In these examples, substitution by synonyms do not work (*i.e.*, not idiomatic). “gaalI denaa” and “gaalI pradan karanA” are synonymous, but not substitutable, because of requirement of idiomaticity. Similarly “lifelong” cannot substitute “life” in “life imprisonment”. The lexemes are very fixed in such expressions.

For multiwords arising from fixity of structure, an example is:

E9: The old man *kicked the bucket* (in the sense of *dying*)

But not

E10: *the bucket was kicked by the old man (to express *dying*)

3.1 How to represent multiwords in UW dictionary

Multiwords represent lexical knowledge and must, therefore, be represented in the UW dictionary, which is a universal repository of lexical knowledge. The LW, i.e., the language word will be the complete MW. But the UW has to be constructed making use of English words, and when a conceptual equivalent does not exist in English, transliteration should be used. For example,

[□□□□□ □□□□□] "son of the soil(icl>person>thing)"; (refer to B5 above)

is a linkage that is possible because of the existence of conceptual equivalents in Bangla and English.

However,

[पोटांत चाबणे] "to feel jealous"(icl>feel)"; (refer to K3 above)

requires paraphrasing the Konkani expression for creating the linkage.

The UNL community has sporadically been deliberating the use of complete UNL expressions in the UW dictionary to express mutliwords. For example, the "feel jealous" concept above could be represented as:

[पोटांत चाबणे] "aoj(feel, jealous)(icl>UNL-expression)";

where, *UNL-expression* is an ontological category under all possible *expressions* that represent knowledge. Technically, this seems like an attractive proposition, since the frozen lexeme "feel jealous" needs cumbersome machinery to incorporate gender-person-number-tense-aspect-modality (GNPTAM) in the expression. On the other hand, the use UNL expressions can incorporate GNPTAM as a sub-process in generation.

5. UW dictionary and wordnet

Word Meanings	**Word Forms**				
	F_1	**F_2**	**F_3**	**…**	**F_n**
M_1	(*depend*) $E_{1,1}$	(*bank*) $E_{1,2}$	(rely) $E_{1,3}$		
M_2		*(bank)* $E_{2,2}$		(*embankment*) $E_{2,...}$	
M_3		*(bank)* $E_{3,2}$	$E_{3,3}$		
…				…	
M_m					$E_{m,n}$

Table 1: Wordnet lexical matrix

Wordnet (Fellbaum, 1998) makes use of relational semantics as the instrument of disambiguation. Table 1 explains this. This matrix is called *lexical matrix*. The rows in the lexical matrix are meaning *ids* (a set of numbers). The columns are word forms. Along the rows, entries represent synonymy. Along the columns entries represent polysemy. Thus for the

word “bank” above, the id *M1* denotes the concept of *depend* expressed by the synonyms “rely”, “bank” and “depend”. Similarly the *M2* row stands for the concept of *embankment beside a water body*. The column marked “bank” expresses the polysemy of the word “bank”, showing the senses of *depend* (M1) , *embankment* (M2) and *financial organization* (M3).

This way of representing unambiguously meanings of words is called *relational semantics*, because the instrument of disambiguation is lexical and semantic relations between words (Cruse, 1986). Relational semantics may be contrasted with componential semantics, where word meanings are expressed as combinations of features. For example, if we have the feature set <*furry, carnivorous, heavy, domesticable*>, the concept of *cat* will be expressed by switching on the features *furry, carnivorous* and *domesticable,* while for the concept of tiger the on features will be *furry, carnivorous* and *heavy*. This kind of 0-1 feature vectors can be used to disambiguate senses of a particular word too. For examples, for the two senses of the word “road” (from wordnet[5]):

1. (95) road, route -- (an open way (generally public) for travel or transportation)

2. (2) road -- (a way or means to achieve something; "the road to fame")

The feature *abstract* will be *off* for the first sense and *on* for the second sense.

The main problem with componential semantics, however, is to come up with a correct and complete set of features. Correct and complete set of features is unattainable due to the world of concepts being a continuum and fuzzy. What, for example, are the features of the concept of *kindness*, other than the rather obvious *abstractness*, and how to distinguish it from the concept of *mercy*? Similar is the problem with verbs, adverbs and adjective. If the feature set is not rich enough, many concepts will be indistinguishable. On the other hand, if the feature set is too fine grained, the feature representation will be cumbersome, confusing and storage-inefficient.

This is the reason why, the device of relational semantics seems to have gained popularity as the methodology of disambiguation. All one has to do is to put together a set of synonyms which, by virtue of their being in the same set, called *synset*, *disambiguate one another*. Thus, though the word “house” is ambiguous, the synset *{house, family}* as in “she is from a noble house” expresses with certainty the *family* sense of “house”.

What if a word does not have synonymy, or has synonyms such that the combination still does not express a unique meaning? The synset *{house, home}* is not disambiguated enough, since the set can mean either the physical-structure sense of “house” (“his house was destroyed in the earthquake of 1942”) or the abstract sense of “home”. In such cases other lexical and semantic relations like antonymy, hypernymy, meronymy *etc.* can be used for disambiguation. Thus,

“talk:*hypernymy*:conversation” (e.g., “A heart to heart talk”)
“house:*meronymy*:kitchen” (e.g., “the house needs repair”)
“kind:*antonymy*:cruel” (e.g., “A kind old man”)
“limp:*troponymy*:walk” (e.g., “The injured man is limping”)

show how relations other than synonymy can disambiguate words. “talk” being a kind of conversation (*“hypernymy:*conversation*”*) has a sense different from, say, a speech, as in “I

[5] http://wordnetweb.princeton.edu

heard a good talk on genetic engineering". "kitchen" being part-of (meronymy) "house" denotes the physical-structure sense of "house". "cruel" being in opposition in meaning (antonymy) to "kind" denotes that it is not the same kind of "kind" as in "what kinds of desserts are there?". The construct "*troponymy*: walk" says that a manner of walking is the sense of "limp" here.

3.1 Comparing and contrasting UW dictionary and wordnet

Lexico-semantic relations are highly effective disambiguators. UW dictionary too employs such relations. Relations called *knowledge based relations* are used to represent UWs. These relations are:

equ: synonymy
icl: hypernymy
pof: meronymy
ant: antonymy

Amongst these *icl* is the most frequently used relation. Consider the UW:

"waddle(icl>walk>do,equ>toddle,agt>thing)" {v} "WALK UNSTEADILY" "SMALL CHILDREN TODDLE"

In UNL, *icl* relation ("kind-of") is used for both hypernymy (for nouns) and troponymy (for verbs). The above UW shows the concept hierarchy maintained in the UNL knowledge base:

Waddle>icl>walk>do

which is a hierarchy of concepts ending in "do", forming part of an ontology, *viz.*, *action*. Further certainty in meaning is brought in by synonymy ("equ>toddle"). The construct "agt>thing" reinforces the verb sense of the concept using the instrument of argument frame.

We can compare the above UW with the wordnet entry for Waddle:

toddle, coggle, totter, dodder, paddle, waddle -- (walk unsteadily; "small children toddle")
=> walk -- (use one's feet to advance; advance by steps; "Walk, don't run!")
=> travel, go, move, locomote -- (change location; move, travel, or proceed; "How fast does your new car go?")

The hierarchy is clearly visible as "waddle>walk>travel…".

In general, one sees a very rich, systematic and deep hierarchy in the noun concepts of the wordnet. The UW dictionary can adopt this structure *completely* for finer knowledge representation through the UWs. The verbal concepts in the UW dictionary, on the other hand, are likely to come out as more expressive and organized, since they propose to use argument frames liberally. The argument frame and selectional preference are built into the definition of verbal UWs. In wordnet, on the other hand, sentence frames are given with verbal concepts. But introduction of argument frame and selectional preference would add to the clarity and richness of representation. Take for example:

propagate(icl>pass_on>do, agt>thing, obj>thing)

in the UW dictionary. This UW has the sense of *transmitting from one generation another* ("propagate the characteristics"). In the wordnet, we see:

Sense 1
propagate -- (transmit from one generation to the next; "propagate these characteristics")
**> Somebody ----s something*
**> Somebody ----s something to somebody*

This representation is not uniform and systematic and is difficult to use in programs.

6. Conclusions and future outlook

In this paper we have discussed some of the issues arising in the construction of UW dictionary based on the content and structure of wordnet. Wordnets are language specific lexical knowledge bases, albeit linked amongst one another[6]. UW dictionary, on the other hand, is aimed at being a universal repository of lexical knowledge. Multilinguality poses a challenge on the way to realizing this universality. Socio-cultural, spatial and temporal influences demand transliteration, paraphrasing and other instruments of representation of concepts in the UW dictionary. Multiwords being non-compositional and/or fixed in lexeme and structure too demand imaginative and new ways of representation. One of the possibilities is to have UNL expressions as UWs. In future, one expects deeper study of UWs as vehicles of lexical knowledge and their linkage with linked open data (LOD) containing *DBpedia, Wikipedia, multilingual wordnets, conceptnet, verbnet, framenet, propbank, Hownet* and so on (Bizer et. al., 2009). This will prove beneficial for the world wide UNL enterprise. The UNL community will also need to think about the semantic web compatibility of UW dictionary and UNL expressions.

References:

Bhattacharyya, P. 2012. *Natural Language Processing: A Perspective from Computation in Presence of Ambiguity, Resource Constraint and Multilinguality*,CSI Journal of Computing, Vol. 1, No. 2.

Bizer, C., Heath, T., Berners-Lee, T. 2009. *Linked Data - The Story So Far*. In: IJSWIS, Vol. 5, Issue 3.

Cruse, D. A., 1986. *Lexical Semantics (Cambridge Textbooks in Linguistics)*. Cambridge University Press.

Dave Sachi, Jignashu Parikh and Pushpak Bhattacharyya, *Interlingua Based English Hindi Machine Translation and Language Divergence*, JMT, Volume 17, September, 2002.

[6] http://www.globalwordnet.org/

Fellbaum, C. (ed.), 1998. *Wordnet an Electronic Lexical Database*. MIT Press.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. 2002. *Multi-word expressions: a pain in the neck for NLP*, proceedings of CICLING.

Uchida. H., Zhu, M. and Della Senta, T., 1999. *The UNL, a Gift for the Millenium*. United Nations University Press, Tokyo.