Some Lexical Issues of UNL

Igor Boguslavsky

Institute for Information Transmission Problems, Russian Academy of Sciences
19, Bolshoj Karetnyj, 101447, Moscow, Russia
bogus@iitp.ru

Abstract. The Universal Networking Language (UNL) developed by Dr. H. Uchida at the Institute for Advanced Studies of the United Nations University is a meaning representation language designed for multi-lingual communication in electronic networks, information retrieval, summarization and other applications. We discuss several features of this language relevant for correct meaning representation and multi-lingual generation and make some proposals aiming at increasing its efficiency.

1 UNL Approach to the Lexicon

The Universal Networking Language (UNL) developed by Dr. H. Uchida at the Institute for Advanced Studies of the United Nations University is a meaning representation language designed for multi-lingual communication in electronic networks, information retrieval, summarization and other applications.

Formally, a UNL expression is an oriented hypergraph that corresponds to a natural language sentence in the amount of information conveyed. The arcs of the graph are interpreted as semantic relations of the types agent, object, time, reason, etc. The nodes of the graph can be simple or compound. Simple nodes are special units, the so-called Universal Words (UWs) which denote a concept or a set of concepts. A compound node (hypernode) consists of several simple or compound nodes connected by semantic relations.

In addition to propositional content ("who did what to whom"), UNL expressions are intended to capture pragmatic information such as focus, reference, speaker's attitudes and intentions, speech acts, and other types of information. This information is rendered by means of attributes attached to the nodes.

After 6 years of the UNL project development, it is possible to take stock of what has been achieved and what remains to be done. In this presentation, I am going to concentrate on one of the central problems with which any artificial language is faced if it is designed to represent meaning across different natural languages. It is a problem of the language vocabulary.

I would like to single out three distinctive features of the UNL dictionary organization.

1. **Flexibility.** There is no fixed set of semantic units. There is only a basic semantic vocabulary that serves as a building material for free construction of derivative

Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 101–108.

lexical units with the help of semantic restrictions. This makes it possible to balance to some extent the non-isomorphism of lexical meanings in different languages.

2. **Bottom-up approach.** The UNL dictionary consisting of Universal Words is not constructed a priori, top-down. Since it should contain lexical meanings specific to different languages, it grows in an inductive way. It receives contributions from all working languages. Due to this, one can expect that linguistic and cultural specificity of different languages will be represented more fully and more adequately than it would be possible under the top-down approach.
3. **Knowledge base.** As the UNL dictionary comprises unique semantic complexes lexicalized in different natural languages, we are facing the task of bridging the gap between them. It is supposed to be done by means of the Knowledge Base – a network of UNL lexical units connected by different semantic relations. Special navigation routines will be developed that will help to find the closest analogue to a lexical meaning not represented in the given language.

There are, however, some circumstances that impede full realization of these features, at least at the moment. Inductive storing of UWs from different languages is a good idea, but this process should be well organized. If a specific UW that is not self-evident is introduced to the UNL dictionary, it should necessarily be supplied at least by an informal comment to make it understandable to other users. Lucidity and easy interpretability of UWs is a goal at which all the developers of the UNL dictionary should aim.

Below, I am going to discuss in more detail two problems that have not so far received sufficient attention in UNL: the argument frames and lexical collocations.

2 Argument Frames

The need to introduce the information on the arguments does not seem to require justification. Any meaning representation language should have an ability to draw a distinction between the argument and non-argument links of predicates. In the UNL expressions, semantic links between the UWs are represented by means of UNL semantic relations. UNL disposes of an inventory of relations which, according to the latest specification, contains 41 items. Here are some examples of the UNL relations:

agt – agent (*John runs*),
obj – object (*read a book, A tree grows*),
ben – beneficiary (*He did not do anything for her*),
cag – co-agent (*I live with him*),
cob – co-object (*He fell into the river with the car*),
aoj – a thing which is in a certain state or is ascribed a property (*I love Mary; my brother is a student*).
dur – duration (*He worked nine hours*),
fmt – a range between two things (*He worked from Monday till Sunday*),
gol – final state (*turn red*),

ins – instrument (*observe with the telescope*),
met – method or means (*separate by cutting*),
pos – possession (*John's mother*),
rsn – reason (*They quarrel because of money*).

It is well known that for correct generation it is essential to know the argument structure of the predicates and the way each argument is expressed in the sentence. The UNL dictionary does not contain explicit information on the argument structure. According to the UW manual, the restrictions which should be included in the UW definitions are not meant for this purpose. As the UNL relations roughly correspond to semantic roles, it is supposed that each argument can be reliably identified based on its semantic role. However, this is not the case. Numerous attempts to construct a set of semantic relations, made over the last decades, showed that only a part of the relations between the words can be unambiguously interpreted in terms of semantic roles. In many cases this interpretation is largely arbitrary. This could not be a problem for the puproses of generation, if it were possible to assign semantic roles in a consistent way. Unfortunately, in practice it is hardly possible, especially when it is done by different people trained in different frameworks and working in different countries. The UNL texts compiled by the UNL project participants from 14 countries over the last years abound in mismatches in the representation of the same or very similar phenomena. Not surprisingly, most of them concern the representation of argument relations. For example, the phrase *base on respect* was interpreted by one team by means of the locative relation (lpl) and by another team by means of the comparative relation (bas), *freedom for all* was described with the purpose relation (pur) and with the beneficiary relation (ben), *bottleneck for the flow of information* received two labels – purpose (pur) and object (obj). Very often, the interpretation of a phrase in the corpus was motivated by the surface form rather than by its meaning. A typical example is *relations among nations* which was described by means of the locative relation obviously under the influence of the literal meaning of *among*. However, nations are by no means the place where relations occur. Rather, nations are participants of the "relations" situation and therefore are more likely to be objects (obj).

Sometimes the motivation behind the use of certain relations may be difficult to understand (at least, this is the case for the author of this paper). For example, in one of the sentences of the corpus, the argument structure of the verb *prevent* was presented as follows:

(1) *Nothing* (obj) *prevents members* (ben) *from discussing* (gol) *this problem*.

In our opinion, these problems are rooted not so much in the erroneous use of relations as in the fundamental impossibility of a consistent interpretation of all argument relations in terms of a small number of semantic roles.

What could one do to avoid the mismatches?

First, one could renounce using semantic roles in cases in which they are not obvious and replace them by semantically uninterpreted relations (subject, first object, second object, etc.). In this case, sentence (1) will receive a more transparent representation:

(2) *Nothing* (subject) *prevents members* (1 object) *from discussing* (2 object) *this problem.*

Obviously, it will be in many cases easier for those who write UNL expressions to develop a common approach to deciding which argument is the first object and which is the second than a common approach to finding appropriate semantic roles for them.

Second, one could accept the proposal of the French team and assign special markers to the case relations when they attach arguments (for example, @A would correspond to the first argument, @B – to the second, etc.). In this case, sentence (1) would be represented as:

(3) *Nothing* (obj.@A) *prevents members* (ben.@B) *from discussing* (gol.@C) *this problem.*

This would certainly reduce the area of uncertainty, but not eliminate it completely. To be able to interpret representation (3), the deconverter should know in advance the argument frame of the UW *prevent*. Otherwise, the uniformity of interpretation will still not be ensured. The only way to eradicate any ground for discordance between different users of the UNL language is to LIST ALL THE ARGUMENT STRUCTURES IN THE UNL DICTIONARY.

To incorporate this proposal, one need not introduce to the dictionary format any new possibilities: the existing apparatus of restrictions is quite sufficient. The only – but very serious – problem is to acknowledge that the argument frame should be explicitly and systematically specified in the UWs. If this is done, then one could keep using semantic roles in all the cases. For example, the word *bottleneck* (in the meaning of an obstacle) can receive the information that its syntactic object (*for something*) has the semantic role "pur" (or any other role which seems appropriate to the lexicographer). If every predicate is supplied with this information in the UNL dictionary, the discordance of opinion between different UNL users will become their private concern and the uniform treatment of the UNL relations in the most controversial zone – that of the argument relations – will be fully assured.

It should be emphasized however that in a general case the marking of the argument frame in a UW is not sufficient either. In some cases the same relation can attach to a UW both an argument and a free adjunct. For example, emotional states (of the type *be afraid, be surprised, be angry*, etc.) have an argument denoting a cause of the state. In sentence (4)

(4) *She is afraid to go out alone at night*

going out alone at night is what makes her to be in the state of fear. Therefore, relation "rsn" between *afraid* and *go out alone at night* is appropriate. On the other hand, *afraid* can have a non-argument cause, as in (5):

(5) *She is afraid (to go out alone at night), because this area is not very safe.*

Even if UW "afraid" is assigned a cause as one of the arguments (afraid(rsn>*)), we should know whether or not a "rsn"-link in the UNL expression denotes this argument. A good solution would be to mark the argument relation by a special label, as proposed in (3). Then, (5) will be represented as (6):

(6) rsn.@A(afraid(rsn>*), go out)
rsn(afraid(rsn>*), safe)

3 Lexical Collocations

Lexical collocations pose a serious problem for any language designed for representing meaning. Here are some examples of collocations from English: *give a lecture, come to an agreement, make an impression, set a record, inflict a wound; reject an appeal, lift a blockade, break a code, override a veto; strong tea, weak tea, warm regards, crushing defeat; deeply absorbed, strictly accurate, closely acquainted, sound asleep; affect deeply, anchor firmly, appreciate sincerely.* For simplicity, I will only dwell below on verbal collocations.

One of the problems such collocations raise is as follows. Some of the members of these collocations do not have a full-fledged meaning of their own. For example, the verb *give* in the collocation *give a lecture* does not denote any particular action. Its meaning, or rather its function, is the same as that of *take* in the collocation *take action*, or that of *make* in *make an impression*. The verbs *give, take* and *make* in these collocations are practically completely devoid of any meaning. Still, they have a very definite function – that of a support verb. This function is exactly the same in all the three cases, and nevertheless the verbs are by no means interchangeable. One cannot say **take an impression, *give action* or **make a lecture.* Moreover, this function is not only performed by different verbs with respect to different nouns. Very often, similar nouns in different languages require different verbs. For example, in Russian a lecture is not given but read, an action is not taken but accomplished, an impression is not made but executed.

How should these phenomena be treated in UNL? In particular, what UWs should be used for support verbs? The current practice suggests that UWs should be constructed on the basis of the source languages. Each language center should produce UWs for the words of its language, without any regard to other languages or any general considerations. A UNL expression and the UWs it consists of are considered adequate if they allow generating a satisfactory text in the same language they originated from. To what extent is this approach applicable to lexical collocations?

To answer this question, we will consider a concrete example. Suppose we have to convert to UNL Russian sentences with the meaning (7), (8), (9) or (10):

(7) *They began the war.*
(8) *We began the battle.*
(9) *The army suffered heavy losses.*
(10) *He took a shower.*

The problem is that in these contexts Russian uses quite different verbs than English. In Russian, correct sentences would be:

(7a) *They undid (razvjazali) the war.*
(8a) *We tied up (zavjazali) the battle.*
(9a) *The army carried (ponesla) heavy losses.*
(10a) *He received (prinjal) a shower.*

If UWs for support verbs in sentences (7a) – (10a) are constructed on the basis of Russian, they would look as follows: “undo(obj>war)”, “tie up(obj>battle)”, “carry(obj>loss)”, and “receive(obj>shower)”. These UWs will allow the Russian deconverter to produce perfect Russian sentences (7a) – (10a). In this case, the condition

for adequacy mentioned above is met. Still, I would not consider UNL expressions based on these UWs adequate. They are produced without any regard for anything except the needs of Russian deconversion and are not fit for other purposes. In particular, these UWs are incomprehensible for anybody except Russians and it is doubtful that any other deconverter will be able to produce acceptable results from them. UWs originating from English will probably look like "take(obj>shower)", "begin(obj>thing)", "suffer(obj>loss)". To generate English sentences (7) – (10) from the UNL expressions constructed on the basis of (7a) – (10a), one would need to somehow ensure the equivalence of UWs "carry(obj>loss)" and "suffer(obj>loss)" in the Knowledge Base. This does not seem to be a natural and easy thing to do. Therefore, UWs for support verbs should not be constructed based on the lexical items of the source language.

Another possibility would be to make use of the co-occurrence properties of English lexical items. UNL vocabulary employs English words as labels for UWs and their meanings – as building blocks for UNL concepts which can be to a certain extent modified by means of restrictions. If lexical labels and meanings of UWs have been borrowed from English, their combinatorial properties can also be determined by the properties of corresponding English words. In this case, UWs and UNL expressions for sentences (7a) – (10a) will be identical to those for (7) – (10).

The advantage of this solution is obvious: since knowledge of English is indispensable for all the developers of X-to-UNL dictionaries, they can be sure that UWs for support verbs they produce are understandable and predictable. This solution has also drawbacks.

First, the inventories of support verbs in different languages are different. Therefore, we will often be faced with gaps in the lexical system of English and find no equivalent for a verb we need. Second, support verbs are bad candidates for the status of UWs. They do not denote any concept. Different support verbs often do not differ in meaning but only in their co-occurrence properties. It seems unreasonable to have different UWs to represent *take* (in *take action*), *make* (in *make an impression*) and *give* (in *give a lecture*), since the difference between these words is not semantic but only combinatorial. This difference should not be preserved in a meaning representation language.

The best solution would be to abstract from asemantic lexical peculiarities of support verbs and adopt a language-independent representation of these phenomena. Theoretical semantics and lexicography have long ago suggested a principled approach to the whole area of lexical collocations. It is the well-known theory of lexical functions by I. Mel'čuk implemented in the Explanatory combinatorial dictionaries of Russian and French (Mel'čuk 1974; Mel'čuk & Zholkovsky 1984; Mel'čuk *et al*. 1984, 1988, 1992, 1999). Possible use of lexical functions in NLP is discussed in (Apresjan *et al*. (in print)). Briefly, the idea of lexical functions is as follows. For more details, the reader is referred to the works mentioned above.

A prototypical lexical function (LF) is a general semantic relation R obtaining between the argument lexeme X (the keyword) and some other lexeme Y which is the value of R with regard to X (by a lexeme in this context we mean a word in one of its lexical meanings or some other lexical unit, such as a set expression). Sometimes Y is represented by a set of synonymous lexemes Y_1, Y_2, …, Y_n, all of them being the val-

ues of the given LF R with regard to X; e. g., MAGN (*desire*) = *strong / keen / intense / fervent / ardent / overwhelming.*

There are two types of LFs – paradigmatic (substitutes) and syntagmatic (collocates, or, in Mel'čuk's terms, parameters).

A substitute LF is a semantic relation R between X and Y such that Y may replace X in the given utterance without substantially changing its meaning, although some regular changes in the syntactic structure of the utterance may be required. Examples are such semantic relations as synonyms, antonyms, converse terms, various types of syntactic derivatives and the like.

A collocate LF is a semantic relation R between X and Y such that X and Y may form a syntactic collocation, with Y syntactically subordinating X or vice versa. R itself is a very general meaning which can be expressed by many different lexemes of the given language, the choice among them being determined not only by the nature of R, but also by the keyword with regard to which this general meaning is expressed. Typical examples of collocate LFs are such adjectival LFs as MAGN = 'a high degree of what is denoted by X', BON = 'good', VER = 'such as should be' and also support verbs of the OPER/FUNC family. Examples of the latter are OPER1 = 'to do, experience or have that which is denoted by keyword X (a support verb which takes the first argument of X as its grammatical subject and X itself as the principal complement)'; OPER2 = 'to undergo that which is denoted by keyword X (a support verb which takes the second argument of X as its grammatical subject and X itself as the principal complement)'; FUNC1 = 'to originate from (a support verb which takes X as its grammatical subject and the first argument of X as the principal complement)'; FUNC2 = 'to bear upon or concern (a support verb which takes X as its grammatical subject and the second argument of X as the principal complement)'.

If used in UNL, lexical functions will ensure a consistent, exhaustive and language-independent representation of support verbs and all other types of restricted lexical co-occurrence. For example, English and Russian support verbs we discussed above – *take* (*a decision, a shower*), *make* (*an impression*), *give* (*a lecture*), *suffer* (*losses*), *prinimat'* (*reshenie* 'decision', *dush* 'shower'), *proizvodit'* (*vpechatlenie* 'impression'), *chitat'* (*lekciju* 'lecture'), *nesti* (*poteri* 'losses') – are correlates of the same lexical function – OPER1.

Being abstract and completely language-independent, lexical functions are devoid of all the drawbacks discussed above and can serve as an optimal solution to the problem of representation of the lexical collocations in UNL.

Acknowledgements This work has been supported by the Russian Foundation of Basic Research (grants Nos. 01-07-90405 and 04-06-80148).

References

Apresjan Ju., I. Boguslavsky, L. Iomdin, L. Tsinman (in print). Lexical function collocations in NLP.

Mel'čuk I. A., 1974. Opyt teorii lingvisticheskix modelej "Smysl – Tekst" [A Theory of Meaning – Text Linguistic Models"]. Moscow, Nauka, 314 p.

Mel’čuk I. A., Zholkovskij A.K., 1984. Tolkovo-kombinatornyj slovar' sovremennogo russkogo jazyka. [An Explanatory Combinatorial Dictionary of the Contemporary Russian Language] Wiener Slawistischer Almanach, Sonderband 14, 992 p.

Mel'čuk I., Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Adèle Lessard, 1984. Dictionnaire explicatif et combinatoire du français contemporain, Recherches lexico-sémantiques I. Les Presses de l'Université de Montréal.

Mel'čuk I., Nadia Arbatchewsky-Jumarie, Louise Dagenais, Léo Elnitsky, Lidija Iordanskaja, Marie-Noëlle Lefebvre, Suzanne Mantha, 1988. Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques II. Les Presses de l'Université de Montréal.

Mel'čuk I., Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Suzanne Mantha, 1992. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III.* Les Presses de l'Université de Montréal, 1992.

Mel’čuk I., Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Suzanne Mantha et Alain Polguère, 1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV*. Montréal.