Информационные процессы, Том 8, № 1, стр. 84-97

============== UNIVERSAL NETWORKING LANGUAGE =============

UNL Graph Structure

V.G. Dikonov

Институт проблем передачи информации, Российская академия наук, Москва, Россия
Поступила в редколлегию 21.03.2008

This work was supported by a grant (No. 04-06-80148) from the Russian Foundation of Basic Research, whose assistance is gratefully acknowledged.

Abstract—This article proposes two improvements over the current UNL graph building guidelines in order to make the logical propositional structure of the graphs more explicit. One is a way to make complex UNL graphs easier to read and edit at low cost and the other is a set of methods and devices to treat coreferential and anaphoric relations in UNL.

1. INTRODUCTION

UNL is an artificial semantic language designed with the goals of multilingual Internet-based information exchange and storage in mind. A text written in UNL is a semantic hypegraph and the core idea of the UNL project is that it is possible to represent the meaning of any human text in this way.

UNL graphs consist of nodes connected with directional arcs. The nodes are filled with «universal words», which represent concepts, or other graphs. The arcs reflect semantic role relations between nodes. Nodes can also have attributes representing various grammatical categories and additional information. The dictionary of universal words is extendable and potentially infinite because it should be able to accommodate all concepts associated with all words of all languages, but the sets of attributes and relations are strictly limited and constant. It makes UNL a simple but very rich and versatile language.

Documents in any human language can be converted into UNL code either manually, which is not practical, or with the help of interactive systems based on deep automatic text analysis. The writer can correct the result by answering questions asked by the system and choosing the correct interpretation if needed. UNL code can also be read and edited directly using special UNL editors capable of visual presentation of the semantic graphs. When someone wants to read a UNL document, the code will be automatically converted into the language chosen by the reader. This process is called deconversion.

UNL can mitigate or avoid translation quality problems that plague conventional machine translation (MT) because it has much less inherent ambiguity than natural languages. As an intermediary language supported by multiple MT systems UNL provides many possible translation pairs and offers great economy of effort while adding new languages. The future UNL-based information infrastructure can significantly enhance access to knowledge by lifting the language barrier and providing very precise and noise free multilingual search facilities.

The UNL project was started in 1996 by Dr. Hiroshi Uchida at the UN Advanced Studies Institute. There are national UNL groups and researchers in many countries, including Russia, France, Spain, Egypt, Japan, China and India, that develop UNL-enabled MT systems for their languages.

2. GOALS

During the course of UNL development a lot of efforts were dedicated towards proper handling of basic sentences while larger units such as complex poly-propositional sentences and text as a whole receive much less attention. It can make the resulting UNL standard less usable than it could be. UNL graphs are meant to be reviewed and edited by people but large graphs are quite awkward to deal with at present. Reading a tangled graph with more than 7-8 nodes can be positively frustrating, especially if the original sentence is written in an unfamiliar foreign language. It is easier for people to trace links and understand a structure if the number of its elements is small. At the same time short elementary sentences are not the dominant type in the scientific, technical and official texts that are most likely to be UNL-encoded. In fact, such texts often contain a lot of extremely complex sentences that produce extremely complicated graphs. Therefore it is a good idea to find a way to make graphs of complex sentences easier to read and understand even for people with no extensive linguistic training. This is goal #1. The best way to achieve this goal is to divide a complex graph into smaller, more manageable parts. Such segmentation has to be very regular and follow a simple rule, yet the graph segments should represent intuitively separable logical parts of the source sentence. This is goal #2. The whole approach should be easy enough to implement without fundamental changes in the UNL formats. This is goal #3. Finally, we shall need a way to establish referential relations between various parts of the graph. This may be considered goal #4.

3. SCOPES IN UNL

Before we go any further we need to decide what means are appropriate for UNL to divide a complex semantic graph into logical parts. Luckily UNL has a convenient and standard way to mark and isolate parts of a graph – scopes, also known as UNL hypernodes. They are supposed to have a very useful feature – an ability to be "folded" in the visual representation of a graph and hide the plethora of nodes and other scopes within and "unfolded" to show the contents. This allows the reader to view the UNL structure with variable level of detail, which makes scopes a perfect tool in order to fulfill our goals.

It should be noted, however, that there are some open questions and different approaches related to scopes. As a brief outline, scopes can be divided by their use into "semantically loaded" and "convenience" scopes. Semantically loaded scopes are needed where a relation link coming to or from a single node is not equivalent to relation with a whole phrase containing this node. They are obligatory, because their presence or absence affects the meaning of the graph. For example, the phrase *old men and women* is ambiguous and requires a scope to specify whether the adjective *old* relates to both men and women:

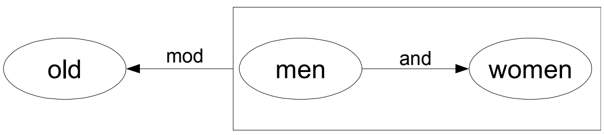


Fig.1 "Both men and women are old"

or solely to men:

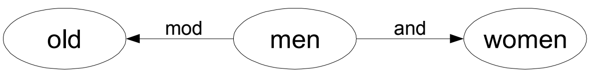


Fig.2 "Men are old but women are not"

While semantically loaded scopes are essential for correct graph interpretation, convenience scopes exist only for the benefit of human readers. They do not change the meaning in any way. Such scopes may be used to group nodes together for the sake of clarity. Additionally, scopes are used to add twin punctuation marks (quotes, brackets, etc.) or text formatting.

There are two approaches to scopes regarding their properties. One considers scopes to be "transparent" and permits nodes within a scope to be arbitrarily connected to any nodes outside of this scope. Such links are known as "cross scope relations". The drawback of this approach is that it undermines the main idea of scope defined by the existing UNL specifications: scopes should behave exactly like nodes. If we try to visualize a graph with cross scope relations and fold the scopes, some of the nodes connected by such relations will be hidden making it impossible to interpret the graph. It shows that transparent scopes cannot always be treated in the same way as nodes.

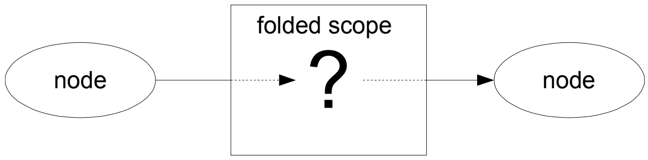


Fig.3 It is not possible to understand a graph with cross scope links while the scope is folded

The other approach treats scopes as monolithic objects. Nodes within "non-transparent" scopes may not have links to any nodes outside. This approach also presupposes that scopes can either incorporate one another or neighbor but never partially overlap. As a result, the graphs become more hierarchical. Such restrictions help to visualize complex graphs because non-transparent scopes can safely be folded to provide a bird's eye view of the structure. It is important that relations will never be shown without the actual nodes they connect as in figure 3 and the graph will always remain understandable.

In general, scopes in UNL are used in two very different ways, which gives reason to make a different division. Some scopes follow the semantic structure. They can be either semantically loaded or convenience scopes but they are always made to reveal the semantic relations between different parts of a sentence. Let us call them semantically motivated. It is beneficial to treat them as non-transparent. At the same time, scopes that are used to show various types of surface mark-up may not be bound to the semantic structure. They exist at a different level and we can call them surface motivated. Unlike semantically motivated scopes, surface motivated ones have to be transparent and should be able to overlap with any other scopes in order not to interfere with the sentence structure and preserve the freedom of formatting. The problem is that UNL does not make this difference and currently supports only one type of scopes.

We shall use semantically motivated, convenience scopes for graph segmentation and try to keep them non-transparent as much as possible.

4. GRAPH SEGMENTATION

At this point we need to choose the basic criteria for graph segmentation. It is not too difficult. Since a UNL graph is a semantic structure, it should be segmented following some natural semantic division. At the semantic level all sentences consist of propositions or elementary "semantic sentences", each one of which expresses exactly one simple situation or fact of the real world. The integral meaning of a sentence is constructed from propositions. This is a well established idea which provides an ideal basis for graph segmentation.

Using scopes to mark propositions might seem easy, but this approach reveals some difficulties. If we try to do it straightforwardly and attempt to mark all propositions in a sentence, we shall see that it is simply not possible in UNL. Let us take an example and try to mark as many propositions as UNL and common sense permit us:

И ускоряя ровный бег	*Accelerating even pace*
как бы в предчувствии погони	*through softly falling snow*
сквозь мягко падающий снег	*as if pre-sensing coming chase*
под синей сеткой мчатся кони.	*the horses under blue cloth run.*
	(A.A. Akhmatova "Ghost")

This Russian sentence is considered to be syntactically simple, but it contains many propositions.

A segmented UNL graph of this verse might be:

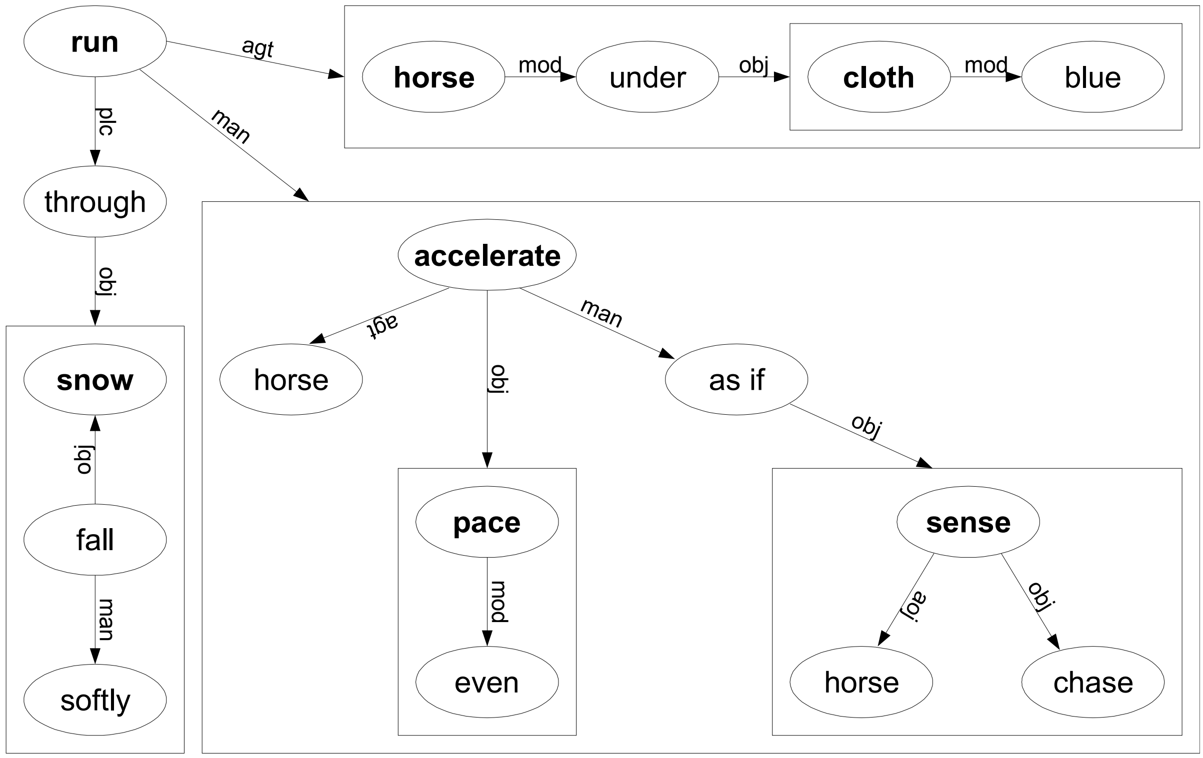


Fig.4 An attempt to mark most obvious propositions
* The UW constraints are omitted for simplicity and entry nodes are shown in bold.

It demonstrates two problems:

Problem 1: The sentence has too many propositions. If we try to put all of them in scopes the graph does not gain in clearness, which goes against our first goal.

Problem 2: Some words belong to more than one proposition at the same time. For example, the node “horse” fills the agentive slots of all verbal UWs except one and has to be part of several scopes. It means that scopes will cross one another, which should not happen in UNL.

However, both problems can be solved in such a way that brings some side benefits.

5. RULES OF SEGMENTATION

The root of the first problem is the fact that every sentence contains a lot of small propositions. The deeper is our semantic analysis, the more propositions we shall find. Therefore we need to be selective and choose which propositions should be marked. All propositions are organized by predicative words, which occupy different positions in a sentence. The most important place in a sentence is that of a verb. Verbs are the most typical vehicle of predicativity. As a result, verbs create the most prominent propositions, which carry the most important parts of the speaker's message and define the structure of the graph. Another point is that verbal propositions are likely to be larger. They include more words or phrases that fill the argument slots of the central predicate. Propositions that are less important to the speaker can be contracted i.e. it is less likely that all arguments of their central words will be expressed in the sentence. It means that scopes of the verbal propositions would typically unite more graph nodes. Besides, verb-centric propositions are very easy to mark, because verbal concepts in UNL are expressed by a distinct class of UWs. Hence the core rule of graph segmentation is:

(1) Enclose all verbal UWs with all their dependents in scopes.

It effectively helps to single out clauses of a complex sentence, infinitive and participial constructions (Assuming that participles are expressed in UNL by verbal UWs). But this formula needs some clarification. First of all, in order to satisfy the principle of compositionality the proposition scopes must include not only first order direct dependents of a verb, but also their daughters.

(2) Follow all dependency links to find all possible dependents and their dependents.

A subtree of all nodes that depend on the chosen root node has to be extracted from the graph. A UNL hypergraph is different from traditional dependency trees because some dependent nodes may have more than one parent and links may ultimately form a circle but it is enough to take into account directionality of relations. Finally, the rule should not require to put verbs without any dependent nodes in scopes, because a scope with only one node inside is useless. Therefore:

(3) If a verbal node has no dependendts, do not make a scope for it alone.

These 3 instructions are enough to fulfill goal #1 to make graphs more clear. The very positive result of the second instruction is that propositions expressed by any subordinate VPs will be put inside the scope of their master proposition while any propositions that do not subordinate each other will be shown side by side. It makes the semantic and syntactic hierarchy of complex sentences instantly visible on screen. The graph of our example verse becomes much more clear with exclusion of previously marked scopes for: *horse under blue cloth, blue cloth, their even pace.*

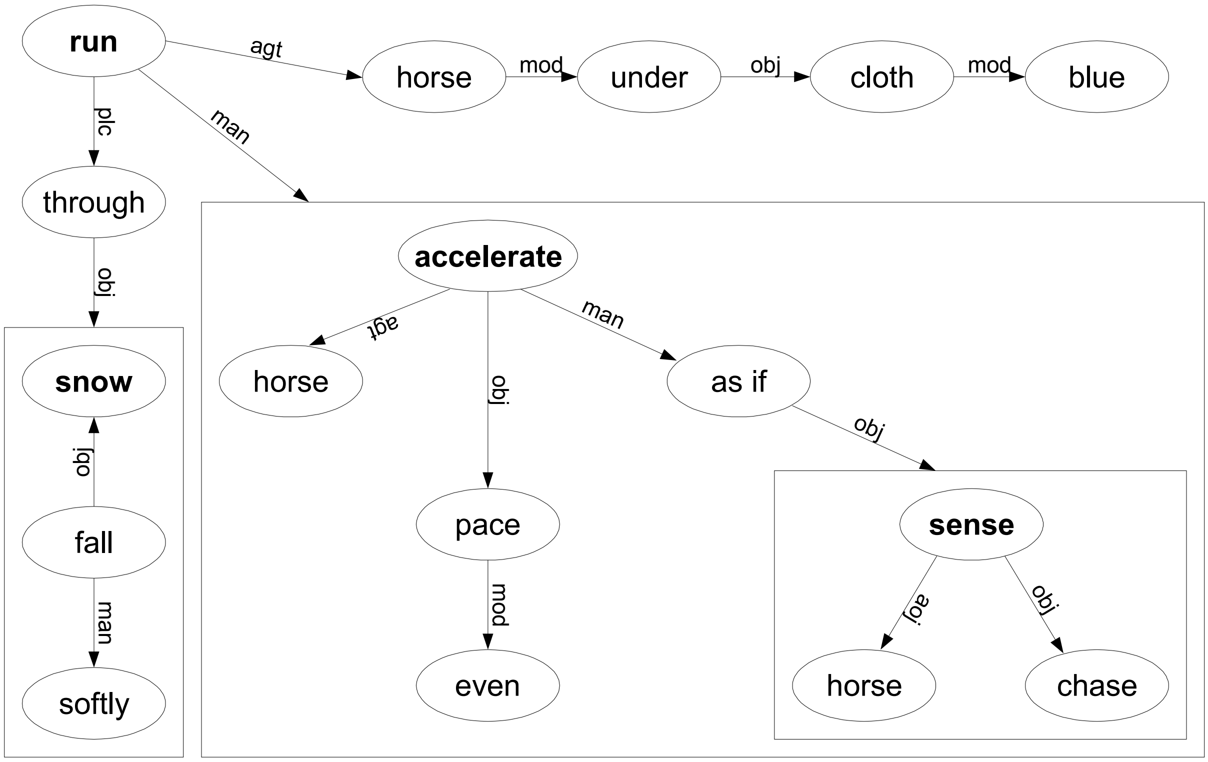


Fig.5 Segmentation based on the verbal propositions only

Now we only have three scopes, which correspond to three participial phrases. They divide the graph into relatively large and well motivated parts. This method is intuitive, because it has firm basis in both semantics and syntax and the required knowledge is taught already in school. It must be noted however that the expected object of segmentation is a UNL graph with links that are reasonably adequate to the structure of the original sentence. It is important to detect all argument relations if we want to extract verbal propositions accurately.

6. SHARED DEPENDENTS

The rules formulated above help to segment a large graph into manageable parts in a simple and logical way. But still there is another already mentioned problem of such segmentation, namely cases when several verbal nodes, that should be in different scopes according to the rules above, share one or more dependent nodes, including ones that fill their argument slots. The most obvious (but not the only) example of this group are sentences with coordinated predicates.

Analysis of an example sentence of this kind can produce a graph which closely replicates its underlying syntactic tree but is semantically incomplete:

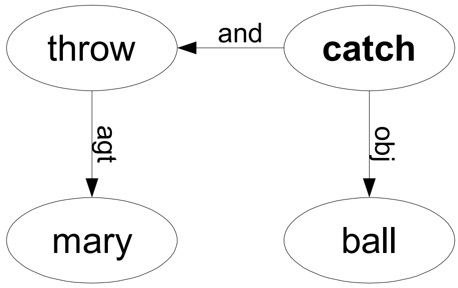


Fig. 6 “Mary throws and catches a ball” – syntax-like structure.

One of the verbs has only a subject and the other has only an object. This is a result of a regular effect, known as “conjunction reduction”. The natural language avoids excessive repetition of the same phrases, but the resulting structure is defective from the semantic point of view. The reduction effect helps to compress spoken sentence and make communication more effective at the physical level of expression without changing the meaning of the sentence. We do understand that “Mary” and “ball” in this sentence are the agent and patient of both actions. However, UNL as a tool to express meaning rather than surface form should encode all argument relations, even if they seem redundant at the other levels of language representation. Restoring the missing links produces a tangle ball of relations:

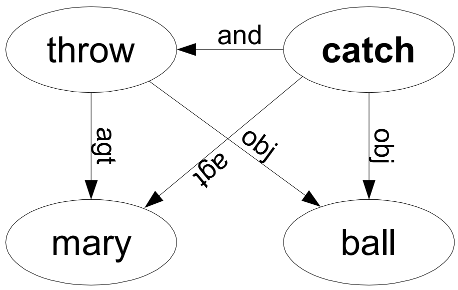


Fig. 7 “Mary throws and catches a ball” after restoring missing argument relations.

This might look good enough for a simple sentence like this example, but with more words the number of relations grows and such graphs become difficult to understand. The proposed segmentation rule makes two subtrees that should be scoped out of this graph, as shown in figure 8.

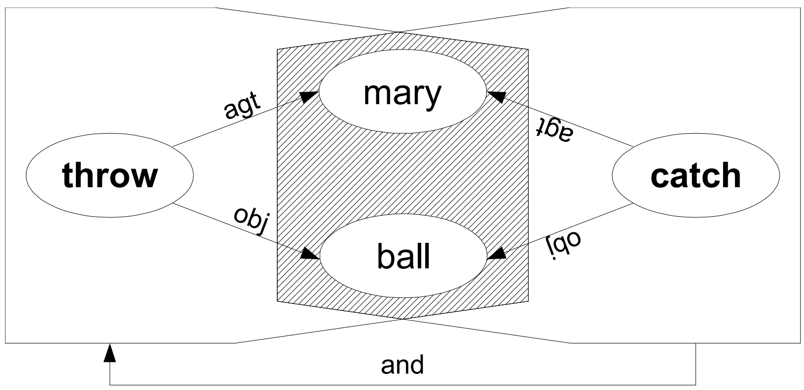


Fig.8. “Mary throws and catches a ball” segmented according to the rule.

The two scopes in the picture correspond to two full propositions, which express both actions performed by Mary. Literally it means that there are two situations where Mary does different things with her ball. However, this example breaks the principle of non-transparency of scopes, which we agreed to follow. The scopes intersect because the nodes “*Mary*” and “*ball*” equally belong to both propositions.

There are different ways to avoid this situation. We can render this sentence in UNL by putting both verbs in the same scope and attaching all arguments to that scope:

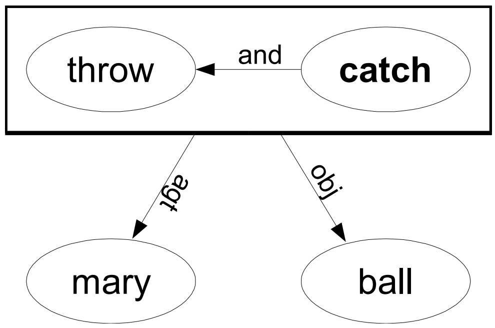


Fig 9. “Mary throws and catches a ball” – the alternative structure.

But the graph in figure 9 has several drawbacks. The argument relations are not attached directly to the verbs and propositions are not shown. One of the consequences is that it is impossible to use this structure if the verbs belong to different classes, e.g. icl>do and icl>be, and require different relations. This is why we believe that expansion of propositions as shown in figure 8 would be preferable if we find another way to avoid overlapping of scopes.

Here we should make a remark. The structure in figure 8 has two propositional scopes while the structure in figure 9 has only one non-propositional semantic scope. It can be formulated as “A + B” and “A” respectively, where the letters denote propositions. This difference has significance and both structures are needed in different situations. Sometimes there is no conjunction reduction and expansion of propositions is not possible. Such cases include:

- integral and continuous actions described by several predicates
  e.g. *I sit and think ≠ “I sit” and “I think”*
- joint and symmetric actions
  e.g. *Mary and Peter quarrel ≠ “Mary quarrels.” and “Peter quarrels.”*

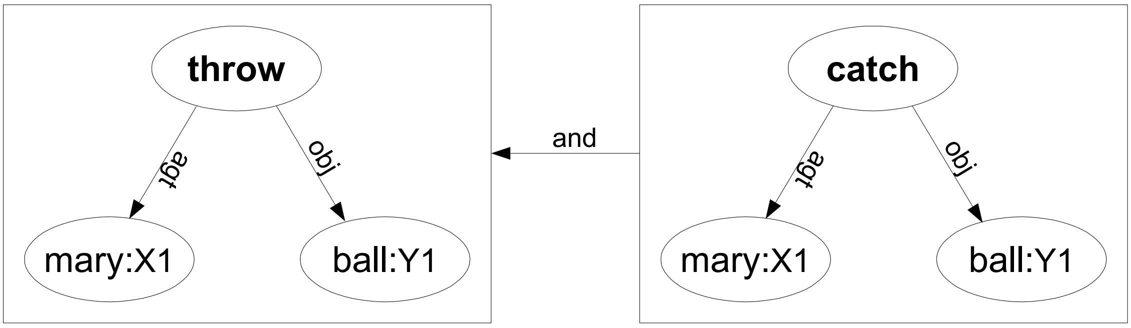
A structure shown in figure 9 is the only appropriate for these examples. At the same time expansion of propositions is the best option for sentences with successive and independent actions like “*He bought a newspaper and went away*”. Finally, some sentences may have dual interpretation. For instance our example may have two slightly different senses: “*Mary [throws and catches] a ball*” (as if playing basketball i.e. both actions are parts of the same continuous process) and “*[Mary throws a ball] and [Mary catches the ball]*” (The actions are momentarily and have no connection other than that Mary is a doer of both). It is also possible to generate two slightly different sentences “*Mary throws and catches a ball*” and “*Mary throws a ball and she catches the ball*” out of the two graphs. It might seem that individual scopes receive some special interpretation here, but this is not true. It is a bad idea to say that any scopes have meaning of their own. However, the meaning of the sentence can be affected by the overall scope structure of the graph even if the inventory of UWs and connecting relations remains the same between two versions. This possibility deserves further exploration.

7. COREFERENCE

As demonstrated in figure 8, the graph segmentation rule creates situations when some nodes have to be members of several scopes in order to show all argument relations and restore full propositions. Therefore, they are drawn twice in the visual graph representation. This brings the discussion about the means of expressing coreference. It is not possible to keep scopes non-transparent and apply the proposed segmentation rule at the same time. If the predicates share dependents and have to be in different scopes and we cannot draw cross-scope links, then we have to invent a method to show all relations between nodes including all their arguments.

A possible answer is node duplication as suggested by Dr. Etienne Blanc in his article “About and Around the French Enconverter and the French Deconverter”. If a node has multiple subordinators located in different scopes, then the node is duplicated. All such nodes have a common identifier to show that they refer to the same thing.

It can be applied to the example graph in figure 8 to separate overlapping scopes:




```
[S:5]
{org}
Мэри бросает и ловит мяч. (Mary throws and catches a ball)
{/org}
{unl}
and(:02,:01)
agt:01(throw.@present.@entry,mary:X1)
obj:01(throw.@present.@entry,ball:Y1)
agt:02(catch.@present.@entry,mary:X1)
obj:02(catch.@present.@entry,ball:Y1)
{/unl}
[/S]
```

Fig. 10 “Mary throws and catches a ball” – node duplication and UNL code.

But if we look at the UNL code it becomes apparent that this is just an oblique way to make cross-scope links without drawing them on screen. The textual representation of the graph remains exactly the same. It means that the relation links from both predicates in this sentence still go to the same UWs and judging by the UNL code the scopes remain intersected. Another point is user-friendliness of this method. What should happen to the visual graph if we need to draw several copies of the duplicated node and it is in turn a scope itself or has multiple dependencies? The graph becomes hard to read because of excessive amounts of multiplicated nodes.

8. REFERENTIAL RELATION

Even though visual duplication of nodes is not always convenient and does not help to keep scopes non-transparent this idea has potential and can be developed further. What is currently expressed by the identifiers of the duplicated UWs is a relation of co-reference. It exists between all instances of the same object in a text and it is not different from the relation between anaphoric pronouns and their antecedents or any other parts of sentence that name the same thing. Referential relations are not peculiar to UNL. They exist in any logically coherent text expressed in any natural language. Correct discovery of such relations is crucial for understanding and acquisition of information contained in the text. Referential relations are not syntactic but semantic. Consequently, they are not bound to the syntactic structures of individual sentences. It makes them intersentennial by nature.

Natural languages have their own devices to show referential relations, including various types of agreement between coreferential words and syntactic patterns. For example, in Russian and many other European languages we have to choose anaphoric pronouns according to the gender class of the coreferential words (anticedents) and maintain agreement in number. Some relations are encoded in the meaning of the sentence e.g. “*Cairo is the capital of Egypt*.” while some have to be inferred logically on the basis of general knowledge and data available earlier in the text, e.g “*The city is situated on the banks of the river Nile*”.

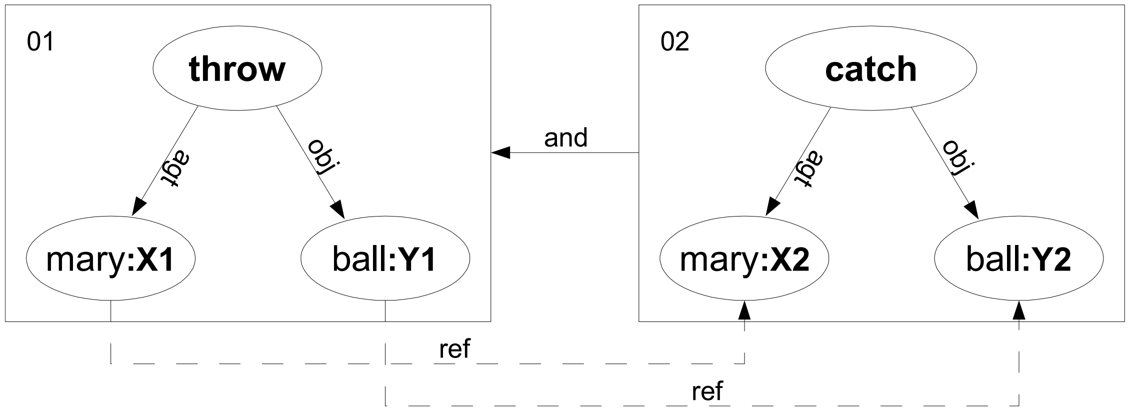
It is possible to express coreference in UNL uniformly as a UNL relation. Let us label it “ref”. The new UNL relation will be used to connect all nodes that refer to the same thing, including any coreferential nodes and anaphoric words with their anticedents. The referential links are directional just like other UNL relations. Their direction is meaningful in the case of anaphoric reference. Such links should go from the anaphoric word to its antecedent to indicate which is where. Coreferential nodes do not require any particular direction, so the "ref" links should follow the intended linear order of the sentence like "and" and "or" .

In order to fulfill its function the referential relation necessarily acquires a unique feature. It should be possible to point it to any node, even located outside of the current sentence graph. Being intersentennial by nature a referential link can cross the scope borders. This seemingly defeats our efforts to make all semantically motivated scopes non-transparent, but due to the nature of the referential relation cross-scope links of this kind do not have the same negative consequences. On the contrary, we can take advantage of this situation and turn their special behavior into a useful instrument. We can still prohibit all cross-scope links except the unavoidable "ref".

The existence of cross-scope referential links does not cause the problem of folded scopes, that was mentioned before (See figure 3), because they do not link arguments to predicates. A UNL graph with referential links that go into or out of the folded scopes will always remain understandable because its structure is defined by other relations. Such referential links are not necessary for understanding the reduced structure created by scope folding and can temporarily be made invisible (See figure 11).

We should also mention the UW identifiers. The new relation does not abolish them even though it takes over their role as the indicator of co-reference. It simply restores their original function as purely technical markers distinguishing UWs that would otherwise be homonymic. Coreferential nodes should be given different IDs now to make them fully independent and prevent intersection of scopes demonstrated in figure 8. This method can be named "UW duplication" as opposed to node duplication.

At this point we can transform our example once again and create a UNL graph which is free from the problems discussed above:




```
[S:6]
{org}
Mary throws and catches a ball
{/org}
{unl}
and(:02,:01)
agt:01(throw.@present.@entry,mary:X1)
obj:01(throw.@present.@entry,ball:Y1)
agt:02(catch.@present.@entry,mary:X2)
obj:02(catch.@present.@entry,ball:Y2)
ref(mary:X1,mary:X2)
ref(ball:Y1,ball:Y2)
{/unl}
[/S]
```

foldable to

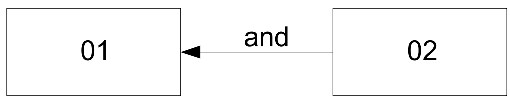


Fig.11 A way to express coreference while keeping the scopes non-transparent.
Folding the scopes creates a reduced graph which remains understandable, though less informative.

The key advantage of this solution is that it helps to keep scopes non-transparent for all argument relations. The replicated UWs become truly separate nodes both in visual and textual representation of UNL. At the same time the graph explicitly and unambiguously shows that all the nodes "*mary*" and "*ball*" refer to the same Mary and same ball.

It is important to note the “and” link between the two proposition scopes. If the scopes are folded it remains the only link to hold the graph together. The “ref” links should be hidden because they connect nodes which become invisible. Besides, any node which is connected with other nodes only by a “ref” link does not have a syntactic place in the sentence. This is another special feature of the referential relation which is based on the nature of coreference. It can be used to show cross sentence links (An example is shown further in Figure 14).

9. ANAPHORIC REFERENCE

Another possible issue created by duplication is demonstrated by the following example:

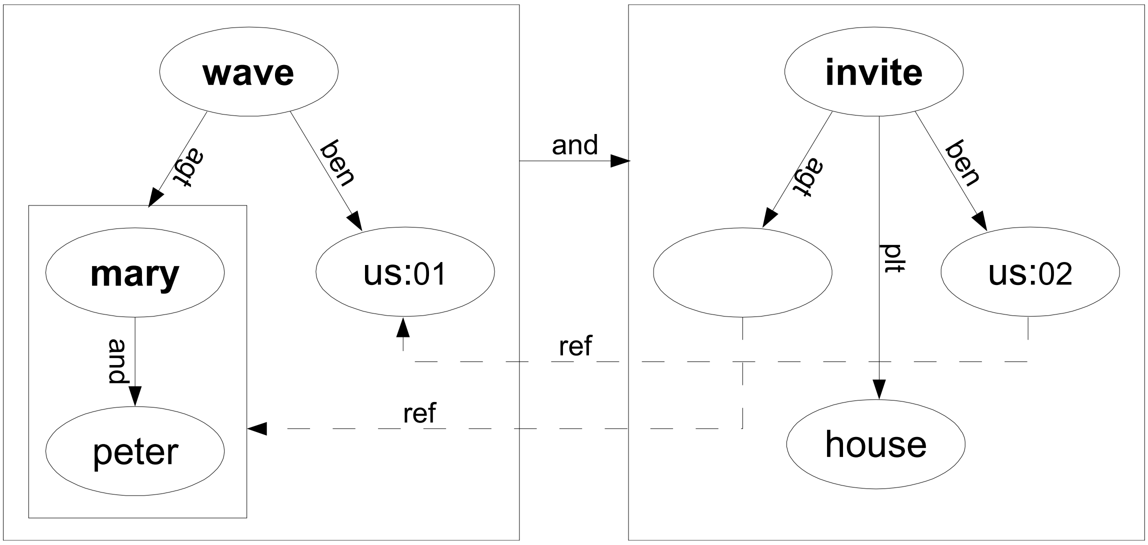


Fig. 12 “Mary and Peter waved to us and invited us to the house”
Copying of scopes or nodes with multiple dependents has to be prevented.

The scope “*Mary and Peter*” in this sentence is an obligatory semantic scope. According to the already stated principles (applied to the UW “*us*”) we have either to make a second copy of this scope or to use some sort of a placeholder for the empty node in figure 12. Having two identical scopes is not desirable because scopes can be unpredictably large. Removing the empty node in the picture is not an option because we need it as the target of the “agt” relation of the verb “*invite*”. Therefore, we need an equivalent of an anaphoric pronoun to be put in this position instead of the scope. However the sentence does not give us such pronoun and finding a suitable candidate is not the task of UNL encoder or reviewer. Therefore, we shall introduce a special anaphoric UW to be used as a substitute for replicated coreferential nodes. Let us name it “*”. The resulting UW is *(icl>thing). It is essentially a generalization of all anaphoric pronouns but unlike pronouns in natural languages it does not enter into any number, gender, etc. agreement governing the choice of pronouns and accepts any anticedents.

It is preferable to use the generic anaphoric pronoun instead of gender-marked English pronouns in all cases of anaphoric reference. A requirement to use English pronouns smuggles in parasitic agreement, which contradicts the idea that UWs represent distinct lexical concepts. All anaphoric pronouns are really one and the same concept. Even though it based on the English lexical inventory UNL is still a separate language. UWs are not divided into the gender classes, so the choice between “he”/“she”/“it” has no basis and is irrelevant for UNL. It would be wrong to enforce it. We shall encode all anaphoric pronouns with the generic pronominal UW as demonstrated in figure 13. This solution frees the encoder of the duty to select a matching pronoun among several English alternatives.

The following graph illustrates various aspects of the referential relation:

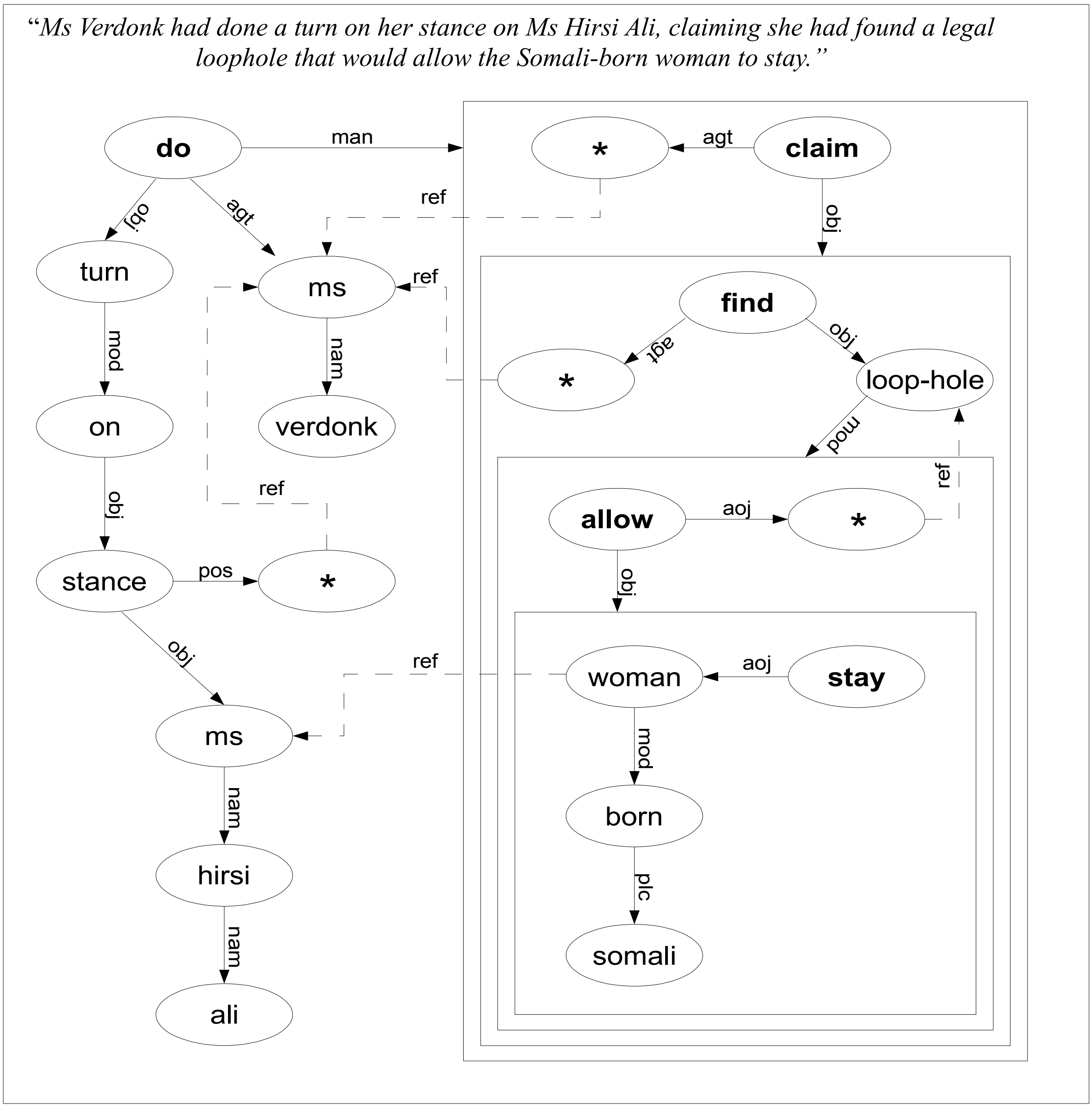


Fig 13. A possible way to express anaphoric and non-anaphoric co-reference in UNL

Some remarks can be made about this graph. First, there are two anaphoras in the English sentence, but the graph contains four. Two anaphoric "*" nodes represent anaphoric pronouns of the source sentence, while two additional ones are the result of UW duplication. All of them are represented by the same UW "*", because they all are substitutes of the nodes connected by the referential links and need the referential relations to be interpreted. Technically all four are anaphoras. The notice that duplicated nodes do not have a corresponding pronoun in the source sentence is not quite correct. Each of the anaphoric nodes fills a required argument slot of some predicate and thus cannot be omitted in the semantic representation. As for the syntax, some syntactic theories state that infinitives and participles have subjects expressed by an invisible or zero pronoun. In this case the two extra anaphoric nodes correspond to such zero pronouns. Since the use of pronouns is highly language specific, it is the generator that must decide, how to render all anaphoric nodes during deconversion. This might improve the percepted fluency of the translated text.

Second, the referential relation is used here to establish a coreferential link between two nominal phrases "*Ms Hirsi Ali*" and "*Somali-born woman*" denoting the same person. This is an example of non-anaphoric coreference, which exists between phrases that refer to the same thing but do not have any direct formal link. These phrases would not have been joined otherwise, and this information would be missing. Such links are valuable for knowledge extraction, if they are present.

Third, the “mod” relation between the node “loop-hole” and the scope “allowing the Somali born woman to stay” is mandatory. The referential relation is neither an argument nor a logical relation. It does not help to assign a syntactic position to the nodes it connects. Without the “mod” link (and the superimposed markup in the form of the [S] tags in the UNL code) the graph, might be interpreted as two sentences “*Ms Verdonk had done a turn on her stance on Ms Hirsi Ali, claiming she had found a legal loophole.*” and “It would allow the Somali-born woman to stay.” with a cross sentence referential link. More about that is explained in the next section.

10. CROSS SENTENCE LINKS

According to the UNL Specification 2005 a UNL text is supposed to be treated as a single giant graph. It is divided into paragraphs and sentences but links between sentences and paragraphs are possible, though the document gives no specific description of how such links should be encoded and processed. A cross-sentence link should go directly to its target in another sentence. Unfortunately, we usually work with isolated sentences. It is a problem, because in order to process a sentence we need information from another sentence that we cannot access. Even if it is not the case, we can still encounter references to extralinguistic or situational knowledge called exophoras. Exophoric words refer to objects that have no linguistic expression in the text. For example the sentence “*Close it!*” does not tell what that “*it*” actually is. Whenever we encounter such cases we need to enrich the informational background during UNL processing to observe agreement required by the target language of generation without injecting the names of extra objects into the generated sentence. None of the existing specifications provides any means to do that, so we might want to propose an amendment to the established formats although such action conflicts with the previously declared goal #3. The proposed way to encode extrasentennial information in UNL graphs is an option. Other proposals in this article do not depend on it.

Following the same principle as with shared dependents we shall create an additional node to represent the referenced extrasentennial object. Such node can be called a “fictitious” or “ghost” node because it does not take any argument roles. A node without argument or logical relations cannot be a part of the sentence, so it disappears during generation. There is a simple rule:

A node that has no other relations within the sentence graph except incoming referential link(s) is fictitious.

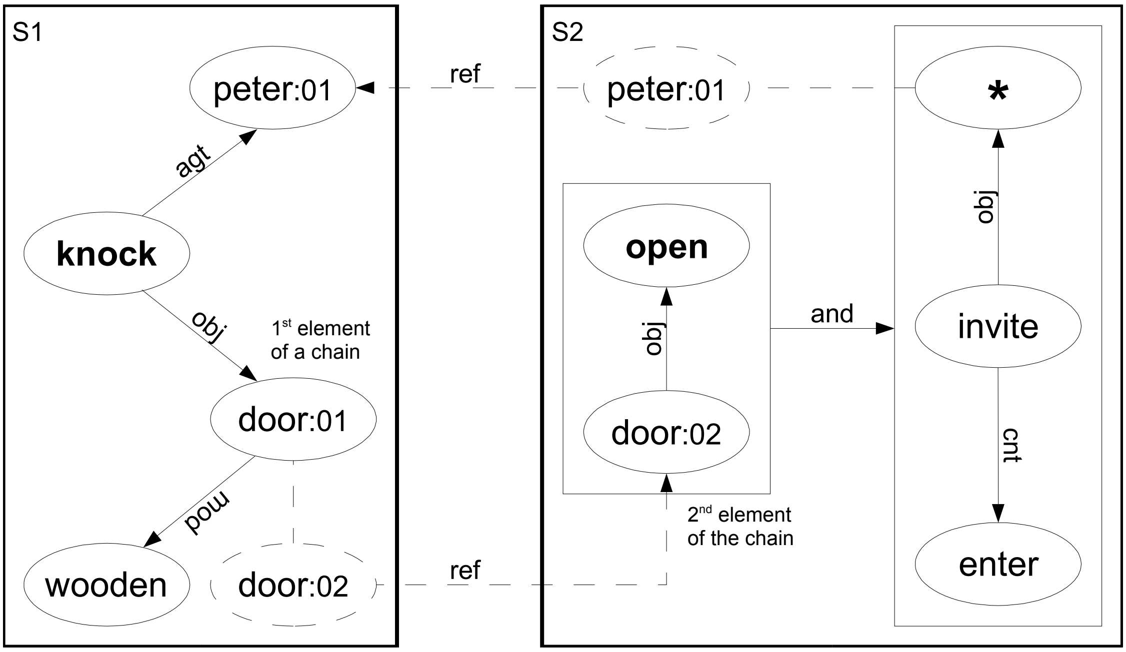
The existence of the “fictitious” nodes is justified by two reasons: First, they provide information required for the choice of an anaphoric pronoun in many languages. To serve this purpose a “fictitious” node should bear the full set of the antecedent's attributes. Second, a fictitious node should uniquely identify the real target of the referential link in another sentence graph. It has to be a copy of the remote UW with the same ID that it has in the other sentence. If the target is a scope, it must be enough to identify it by the scope number. This is the original method of node duplication as suggested by Dr. Etienne Blanc. It was criticized earlier because it creates cross-scope links, but intersentennial referential links are special by nature and can be cross-scope, as explained before. This special feature makes the method of node duplication applicable for intersentennial relations. An example is given in figure 14.

There is an important consequence. In order for the intersentennial node duplication to work, the UW and scope identifiers should be consistent text-wide rather than sentence-wide. Now, if we have to assign unique ID to every node and scope in a whole text we would need more IDs than possible combinations of just two alphanumeric symbols. A possible answer is purely numeric IDs without length constraint. This eliminates the limit of the maximum number of nodes in a text.

11. REFERENTIAL CHAIN

If more than two nodes are connected by the referential relations, they form a referential chain which incorporates all instances of a named entity, even if the names are different. All nodes sequentially connected by referential links should be considered coreferential throughout the text. This opens up a number of possibilities both for UNL deconversion and processing of knowledge stored in the UNL format. For example, it would be extremely useful for correct assignment of articles if the source graph is derived from a language without articles such as Russian and contains no @def and @indef attributes. The first element of a chain corresponds to the first mention of the object and receives the indefinite status for English and languages that use this category in the same way, while the rest represent subsequent references to that object and become definite.

The following picture illustrates a cross-sentence anaphoric referential link and a minimal referential chain:



[S:1]
{org}
Peter knocked at a wooden door.
{/org}
{unl}
agt(knock,**peter:01**)
obj(knock,door:01)
mod(door:01,wooden)
ref(door:01,door:02)
{/unl}
[/S]

[S:2]
{org}
The door opened and he was invited to enter.
{/org}
{unl}
obj:01(open,**door:02**)
obj:02(invite,*)
cnt:02(invite,enter)
and(:01,:02)
ref(*,peter:01)
{/unl}
[/S]

Fig. 14 Cross-sentence referential links.

There are two "fictitious" nodes for cross-sentence links drawn with a dashed line. The node "*door:02*" in the first sentence represents the link to the second member of the referential chain. The node "*peter:01*" in the second sentence is needed to explain the anaphora with the anticedent located in the first sentence.

12. CONCLUSION

This article proposes to put verbal propositions in convenience scopes whenever possible and make referential links explicit in UNL by turning them into UNL relations. The result is two-fold. Scopes make the graphs easier to understand by dividing them into smaller parts and revealing the semantic hierarchy of the text. On the other hand, the proposed formal rule of proposition expansion may result in longer UNL code. However, even if the number of nodes and lines of code may increase, the overall structure becomes simpler, more regular and better organized. Introduction of the referential relation makes it possible to restrict cross-scope links in favor of the non-transparent use of scopes. It also permits to use anaphoras and express non-anaphoric coreference in a regular way preventing the loss of this information in UNL. The core innovations leave the existing UNL format unchanged. The only slight modification of the convention regarding the UW identifiers is strictly optional. Still, it deserves consideration, because it opens an exciting possibility to encode cross-sentence references and referential chains even if sentence graphs are processed in isolation from the whole text. Implementation of the proposals put forward in this article will make UNL better suited for automated knowledge extraction. Expansion of full verbal propositions and referential links might take some load off AI summarization, search and learning tools and can improve their output and performance. This is a valuable result, because UNL is going to be used for different purposes that are not limited by translation alone.

REFERENCES

1. Blanc E. About and Around the French Enconverter and the French Deconverter *Universal Networking Language: Advances in Theory and Application vol.12,* Instituto Politechnico Nacional, Mexico 2005
2. Boguslavsky I., Some Controversial Issues of UNL: Linguistic Aspects. in *Universal Networking Language: Advances in Theory and Application vol.12*, Instituto Politechnico Nacional, Mexico 2005
3. UNL Specifications (version 3.3), http://www.undl.org/unlsys/unl/UNLSpecs33.pdf
4. UNL Specifications 2005, http://www.undl.org/unlsys/unl/unl2005/