English/Russian to UNL Enconverter[1]

Viacheslav Dikonov

Computational Linguistics Lab – IITP RAS
Bolshoy Karetny per. 19, Moscow, 127994, Russia
dikonov@iitp.ru

Abstract

This paper presents a new UNL enconverter - an automatic rule based tool to transform English and Russian text into semantic graphs. It is built on top of the ETAP-3 linguistic processor, which is largely based on the general linguistic framework of the Meaning↔Text theory by Mel'čuk. We describe the principles and operation of the enconverter and highlight its key features. The UNL graphs produced by this system are understood by several UNL deconverters which transform them back into text in different languages. The system represents a step further beyond the level of deep syntactic structure towards building a complete Text→Meaning model.

Keywords

UNL, ETAP, semantic graph, semantic representation, rule based system, interlingua

1 Introduction - What is UNL and UNL enconverter?

UNL (*Universal Networking Language*) is an interlingua based on the semantic graph formalism. It is designed to capture the meaning of text in any natural language with good precision, store it in the computer memory and support generation of equivalent text in any other natural language. The semantic UNL graphs are both machine and human readable and can be post-edited to correct any errors introduced by automatic analysis. A *UNL enconverter* is a software instrument to analyze a natural language text and represent its meaning as a UNL graph. The reverse transformation is performed by *UNL deconverters*.

A UNL graph consists of nodes linked with directional semantic links of several broadly defined types, such as "agent", "partner", "posessor", "place", "purpose" etc. A node may have multiple incoming edges. The nodes may contain either single concepts or other UNL graphs. Nodes of the latter kind are called *hypernodes*.

[1] This work received finacial support from RFBR (Grant 08-06-00367).

The basic lexical units of the UNL language are *concepts* usually representing a sense of a word or an idiom of some natural language (not necessarily English) as distinguished by explanatory dictionaries. There are concepts corresponding to abstract ontological terms as well, e.g. "*thing*", "*do*", "*abstract_thing*". Each concept has at least one[2] label called *UW* (*Universal Word*). Each UW expresses exactly one concept. The idea is that all concepts of all languages should receive unique UWs. If UNL lacks a UW for some concept, it can be added on demand. A UW consists of a headword and a list or constraints clarifying its meaning, e.g. *drink(icl>consume>do,equ>imbibe,agt>person,obj>matter)*. The headwords are usually English words, but words of other languages in Latin transcription may be used too. The constraints are UNL relations linking the concept to other concepts.

The graph nodes can carry special attached marks called *attributes*. UNL attributes are used to encode common grammar categories, subjective standing of the speaker towards the spoken, logical relations, etc. There is a finite set of possible attributes. The names of all UNL attributes start with the sign @, e.g. @past, @unreal, @topic, @polite, @possibility...

UNL graphs resemble both deep syntactic and semantic structures described by Mel'čuk (Mel'čuk, 1974). Table 1 summarizes key differences between UNL and SemStr.

SemStr	**UNL graph**
Semantemes can be decomposed into graphs consisting of elementary semes.	UNL does not decompose its concepts into any smaller units.
Graph edges in SemStr are nameless or numbered. They can be interpreted as semantic links only after consulting the dictionary entries of the predicates.	All graph edges must be labeled with one of 46 predefined semantic types. The resulting links retain their meaning even when the lexical contents of the nodes is masked.
Links are always directed from predicates to objects.	UNL still has the syntactic *mod* (modifier) relation which goes from an object to the modifying predicate. The inverse relation *aoj* (thing with an attribute) is also available.
A regular SemStr should contain only nodes and edges.	UNL graph nodes can have attributes. Some attributes, e.g. modal ones, can be expanded into additional nodes and edges. Another characteristic formal feature is the use of hypernodes.
SemStr should include discourse indicators as « metapredicates ».	Attributes can serve as indicators of topic and emphasis, but there is no established way to indicate theme and rheme. In order to preserve the sequential order of sentences within a text, they are represented as separate numbered graphs.

Table 1: differences between SemStr and UNL graphs

2 What is inside the UNL enconverter?

Our UNL enconverter operates on top of the ETAP-3 syntactic parser and uses the ETAP-3 (Apresjan et. al 1992) software platform, its dictionary and rule formats. It consists of the ETAP UNL dictionary and a set of rules to produce UNL graphs.

[2] It is unrealistic to say that any concept will always have just one associated UW, because there are multiple UNL-related projects. UWs may be created and edited by different people and even after merging all equivalent UWs together in a common dictionary alternative forms still have to be recognized in legacy documents encoded using older versions of UNL.

2.1 UNL Dictionary

The current UNL dictionary in ETAP, which is directly used by our enconverter contains over 82 000 UNL concepts linked with over 47 000 English and 22 000 Russian words. It is a subset of the general UNL dictionary of concepts, expressed in the native ETAP format. It includes only the UWs that have direct association with the words registered in the English combinatorial dictionary of ETAP.

The general UNL dictionary of concepts is larger and is currently under active development. It already includes some data currently unused by ETAP, such as French-UNL dictionary and a large ontology. The structure and principles of the general UNL dictionary are described in (Boguslavsky & Dikonov, 2008). The resource is proposed as a common standard for several research groups, which joined into the so-called “U++ Consortium”, namely the Russian group at IITP (Moscow), the French group at CLIPS-IMAG (Grenoble), the Spanish group at UPM (Madrid) and the Indian group at IIT (Mumbai). The UWs in the common dictionary are also linked with Princeton Wordnet synsets and terms of the SUMO ontology. It is an open resource which includes data from other open resources and our plan is make it available for the public under GPL and CC licenses.

2.2 UNL en-conversion pipeline

In order to convert a text into UNL graphs ETAP splits it into sentences and performs full syntactic parsing of each sentence first. The result is a dependency tree closely following the Mel'čhuk's formalisms. It should be mentioned, that our approach to UNL en-conversion is totally based on dependency trees . The UNL en-conversion process begins after obtaining the syntactic parse. It includes three well defined stages:

- Preliminary markup and light normalization according to UNL rules,
- Lexical disambiguation,
- Semantization of syntactic relations & creation of hypernodes.

Each of these steps is served by dedicated sets of rules and will be described further.

Perhaps, the best way to explain the operation of any natural language processing system is to show how it would process a real sentence step by step. Let us take a random example from a news article about the E.Coli epidemic in Germany (1) “*They said that despite nearly 200 new cases in Germany - the centre of the outbreak - infection rates were dropping.*” and describe all intermediate states and its internal representation in the system.

2.2.1 Syntactic tree

First, ETAP produces a single dependency tree. The tree undergoes normalization, which turns it into the so-called *Normalized syntactic structure* (*NormSS*) used by ETAP for transfer-type translation into other languages.

The UNL enconverter takes English NormSS as its input. If we choose to analyze Russian, the Russian NormSS will be converted into the English NormSS before enconversion. The inevitable losses caused by such conversion are partially offset by storing the default UNL equivalent (the most common sense) of Russian words without direct English equivalents in the special internal field for default translations. Figure 1 shows the initial stage of enconversion.

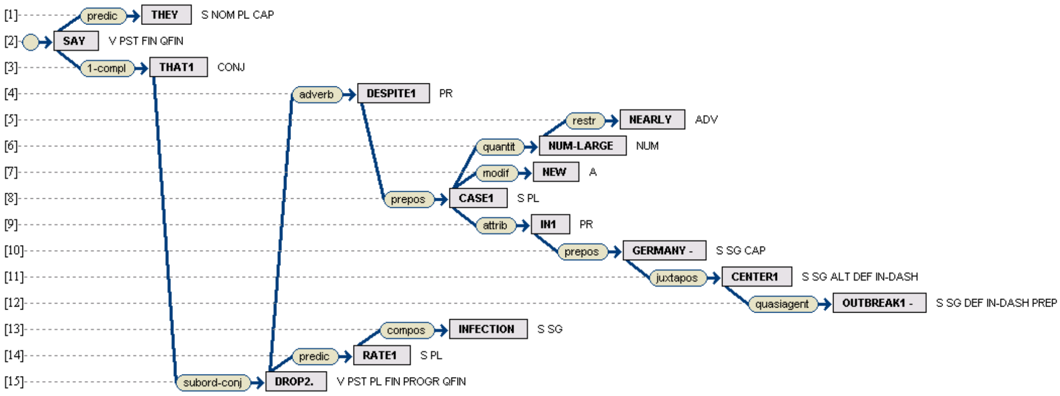

Figure 1: Normalized syntactic structure of example 1 before conversion.

2.3 Preliminary markup

The first step of preliminary markup serves to reduce complexity of further processing and avoid certain errors by inserting UNL attributes that will guide other rules at further stages. The overall structure of the syntactic tree and contents of its nodes usually remain unchanged.

There are several important tasks performed at this stage:

- The system detects if the current sentence contains questions, addresses or imperatives and adds appropriate UNL attributes.
- Tense, aspect and actuality information is analysed and UNL attributes of tense (@past, @present, @future), aspect (@complete, @progress, etc.) and unreality (@unreal) are inserted. If the sentence contain verbs should/would/could/might, special rules are used to determine whether they express tense, subjunctive mood or modality to prevent false interpretations.
- If the sentence contains passive constructions, the UNL attribute @topic is inserted to mark the real object of the action and the passive construction is converted to active by replacing "predic" links with "1-compl".
- Future head nodes of the graph and some of its hypernodes are marked. This rule finds the main predicate and all of its coordinated predicates and inserts special internally used attributes that request (@_HN) or prohibit (@_HNTOP) the creation of hypernodes with the marked predicates as their heads. It prevents putting the whole sentence into a hypernode.
- There are also rules to markup the semantics of participles. The phrases like *"killed people"* and *"drunk people"* are syntactically similar, but the killed are patients of killing while the drunk are agents of drinking. This fact must be reflected in the UNL graph as *killed* $\xrightarrow{obj}$ *people* and *drunk* $\xleftarrow{mod}$ *people*. Likewise, present participles can denote either states or action. For example *"the boy living in Sweden"* produces *live* $\xleftarrow{mod}$ *boy* and *"the boy driving the car"* becomes *drive* $\xrightarrow{agt}$ *boy*.

- Certain anaphoric constructions are detected and a proposed special relation "*ref*" is created between the anaphoric word and its antecedent. Here are some examples: *"Tom with his toy"* (he→Tom) *"Jane and her bag"* (Jane←her) *"Tom and Tom's books"* (Tom^1 = Tom^2) *"Sir, i brought your dinner"/"Your dinner, sir"* (Sir←you) *"This is for you, my dear."* (you→dear). It is impossible to detect any anaphoras which span several sentences because ETAP processes just one sentence at a time and does not store any information about previous sentences.

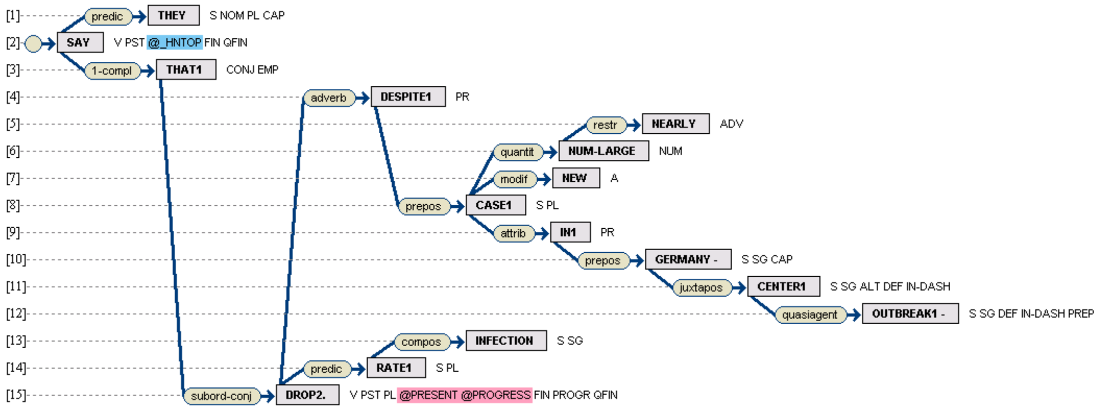

Figure 2: Preliminary markup of example 1.

In our example sentence (Fig.2 above) we see only two changes. The service attribute @_HNTOP (in blue) marks the main predicate of the sentence and prevents it from becoming the starting point of an unnecessary hypernode that would wrap the whole sentence. Other two UNL attributes @present and @progress show that the time of rate dropping was present relative to the moment of speech and this process has not ended by that moment.

2.4 Lexical disambiguation

The goal of this step is to replace all words in the tree with suitable UWs. No syntactic links are changed here and the structure still remains a tree (See Fig.4). Each word can be associated with several UWs representing its different senses. The choice of the correct word sense and UW is important for further processing beause UWs provide access to semantic information in the UNL dictionary. It includes ontological classification and the data necessary to interpret completive syntactic links in terms of UNL relations.

Lexical disambiguation is a classical problem, which has no good solution so far. The ETAP platform offers just three possible ways to do it without turning to external tools. First, all UWs are ranged according to their frequency measured against the Semcor corpus. It helps to improve the number of correct guesses by taking the first – most frequent – sense in the list. Second, there is a system of rules that analyze the immediate tree context and automatically insert the right UW if the context matches a known scheme. Finally, the system has an interactive mode, shown in Fig.3, which allows to ask the user about the true meaning of the words in the analyzed sentence. This method achieves high precision but it is time consuming and tedious for the user.

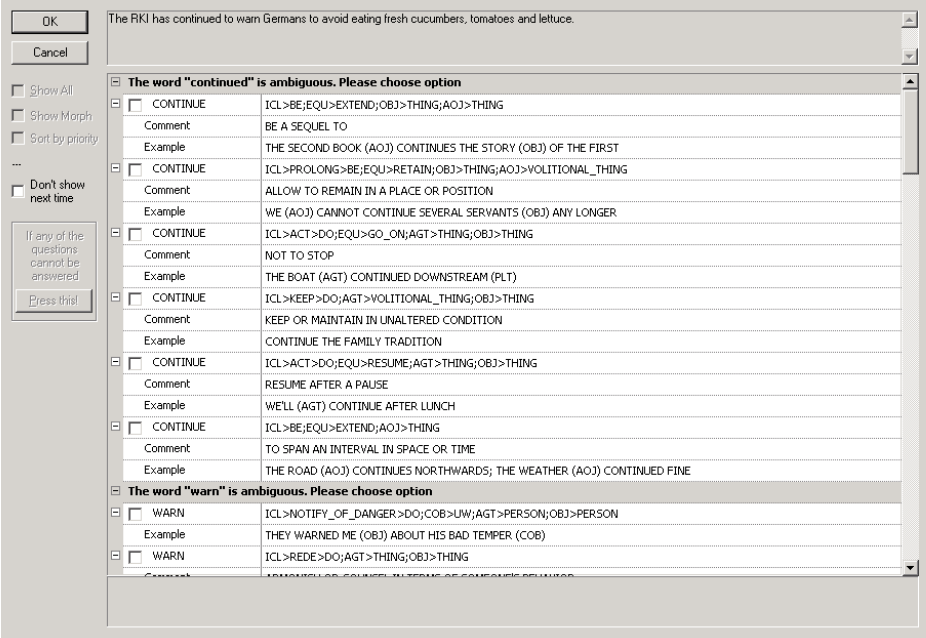

Figure 3: Interactive disambiguation for UNL.

The efficiency of the interactive disambiguation can be improved by changing the user interface so, that the user would see the sentence preannotated with automatically chosen UWs and change only the false choices by indicating one or more suitable ones. Unfortunately, this has not been implemented.

One of the major problems with disambiguation rules is that there must be a set of rules for each word sense describing its possible contexts. Their number could easily go over a million. To mitigate this problem we use a limited set of less than 40 very simple generalized macro rules matching small predefined fragments of the NormSS tree. The fragments may include the parent and 1 – 3 dependent words with their morphosyntactic and semantic features and relations. The actual words, names of relations and features are substituted by variables. The exact values of the variables depend on the currently analyzed word and are supplied by the dictionary. Let us have a look at one of the simplest rules, called UNL-CONV1.15:

REG:UNL-CONV1.15 TRANSLATION DEPENDING ON FEATURES OF A DEPENDENT LEXEME //ВД
N:1 — *Subrule number 1*
CHECK — ***Conditions that must be satisfied in order to apply the rule***
1.1 DOM-EQU(X,,T1,T2)* — *Current node X commands another node * with feature T2 via relation T1.*
2.1 PARAMSYS(DISAMBMODE,DISAMB_LEX,DISAMB_LEXSYNT) — *Do not override human choice in interactive mode.*
2.2 =(X,EMP) — *Skip nodes marked as semantically empty.*
DO — ***Actions performed if all conditions are satisfied (This part is the same for all such rules.)***
1 ZAMUZ:X(UNL,LUNL) — *Replace the word in node X with the UW given by the LUNL variable.*

In plain language it reads as follows: "*If the current word has a syntactic daughter linked by relation T1 and this daughter has syntactic or semantic feature T2, choose the UW specified by LUNL*". The verb "DROP" has the following records in the dictionary:

...
TRAF:UNL-CONV1.15 (reference to the rule template shown above)
T1:3-COMPL,T2:'OBJECT',LUNL:DROP{ICL>MOVE>BE;EQU>DESCEND;PLT>THING;PLF>THING;AOJ>CONCRETE_THING}
T1:3-COMPL,T2:'HUMAN',LUNL:DROP{ICL>UTTER>DO;AGT>PERSON;OBJ>ABSTRACT_THING;REC>PERSON}
T1:PREDIC,T2:'PARAMETER',LUNL:DROP{ICL>CHANGE>OCCUR;OBJ>THING} *"RATES ARE DROPPING"*
...

The last line causes the system to interpret "drop" as "abrupt decreasing" in the context of a parameter. Such rule references come in sequence ranged from more complex rules with more parameters to simpler ones with less parameters. This forms a decision tree that can be compared with sorting of diamonds, where there are many sieves with progressively bigger holes. Smaller gems (better defined contexts) fall through the first ones and only the largest reach the last sieve (the most general context). It is also possible to specify the second best answer for each defined context by setting a special flag "OPT:1". The user can manually ask the system to produce a different version of the graph automatically and the alternative outputs will contain such second answers. If there are no rules or none of them provides an answer, the first entry in the list of UWs associated with the current word will be chosen.

Although this mechanism may seem to be quite old-schoolish, it has been designed with the possibility in mind to use trendy statistical methods to generate massive amounts of such rule entries by analyzing a corpus.

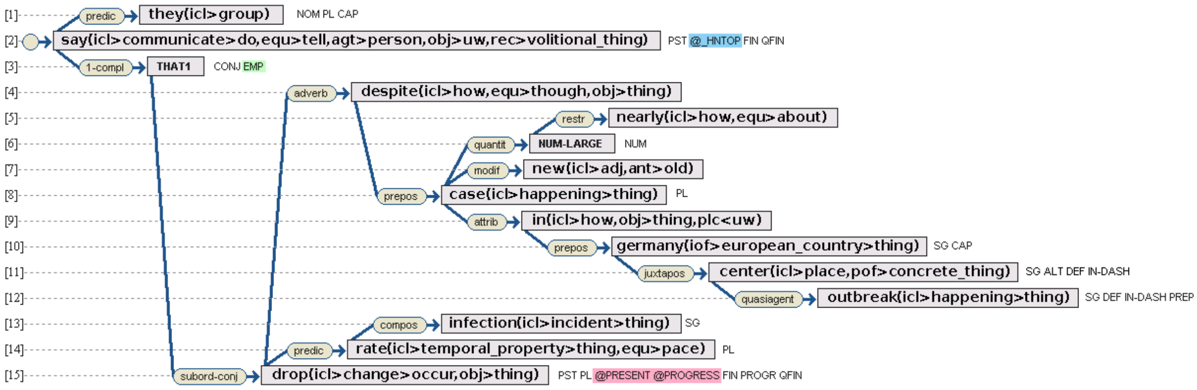

Figure 4: Lexical disambiguation of the example sentence 1.

In Figure 4 there are two words that have not received any UW: the number "200" and "THAT1", which is marked as semantically empty by the EMP attribute. All words with this mark will be deleted at the next stage. Prepositions, such as "IN" in our example, and conjunctions are assigned lexical meanings too.

2.5 Semantic interpretation of syntactic links

This step is dedicated to the actual transformation of the syntactic tree into a true semantic graph (See Fig.5 below).

```
[S:00]
{org:en}
They said that despite nearly 200 new cases in Germany - the centre of the outbreak - infection rates were dropping.
{/org}
{unl}
agt(say(icl>communicate>do,equ>tell,agt>person,obj>uw,rec>volitional_thing).@entry.@past,they(icl>group).@pl)
man:02(drop(icl>change>occur,obj>thing).@entry.@pl.@present.@progress,despite(icl>how,equ>though,obj>thing))
man:02(200,nearly(icl>how,equ>about))
qua:02(case(icl>happening>thing).@pl,200)
mod:02(case(icl>happening>thing).@pl,new(icl>adj,ant>old))
obj:02(despite(icl>how,equ>though,obj>thing),case(icl>happening>thing).@pl)
plc:02(case(icl>happening>thing).@pl,germany(iof>european_country>thing))
obj:01(center(icl>place,pof>concrete_thing).@entry.@def,outbreak(icl>happening>thing).@def)
mod:02(rate(icl>temporal_property>thing,equ>pace).@pl,infection(icl>incident>thing))
obj:02(drop(icl>change>occur,obj>thing).@entry.@pl.@present.@progress,rate(icl>temporal_property>thing,equ>pace).@pl)
cnt:02(germany(iof>european_country>thing),:01.@dash)
obj(say(icl>communicate>do,equ>tell,agt>person,obj>uw,rec>volitional_thing).@entry.@past,:02)
{/unl}
[/S]
```

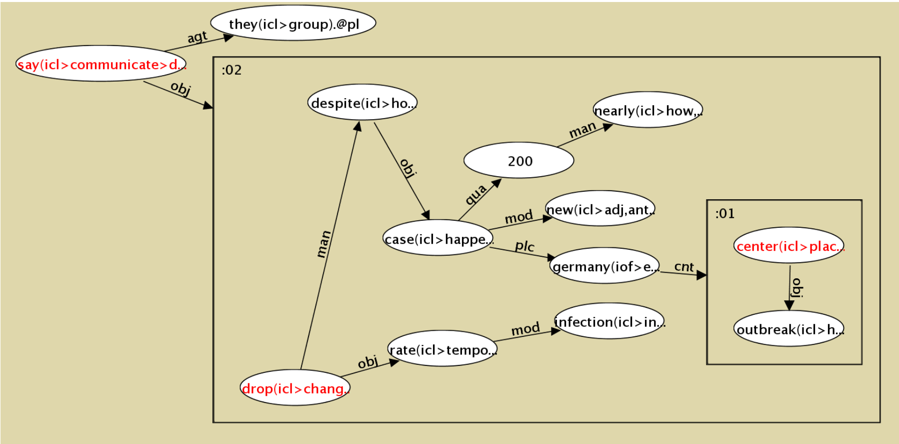

Figure 5: The resulting graph and UNL code of the example sentence 1.

At this stage the structure may cease to be a tree. All syntactic links except the emergency "fictit" link must be replaced by UNL relations. There are two tiers of rules that do it: the macro rules, which perform transformations specific to the chosen UWs, and general rules. First, the macro rules replace 1,2,3,4-completive and other links representing the government patterns of predicates with corresponding semantic relations specific to each UW. The links around prepositions and conjunctions are treated in the same way depending of the chosen meaning. Second, the general rules interpret remaining syntactic links and replace syntactic features with corresponding UNL attributes. Finally, the hypernodes are created where necessary. Figure 5 shows the final result of the conversion. This particular graph contains two hypernodes. Node :02 describes the contents of what has been said and :01 contains the inserted side note about Gemany "*the center of the outbreak*" and holds the attribute @dash to indicate original formatting. The empty word THAT has been deleted and the preposition IN in its spatial meaning was reduced to the equivalent relation "plc" (place). All information about the original word order is erased. The red nodes at the graph's root and in each hypernode are the starting points for graph interpretation.

3 Highlighted features

The UNL enconverter project introduces some important new features and resources both to ETAP and UNL. The most important new features in ETAP were:

- interlingua-based translation and potential support for extra language pairs via UNL deconverters built by other groups,
- support for hypernodes and non-tree graph structures,
- ability to modify the rule behavior depending on the system's operating mode,
- a new dictionary with extra semantic information.

Most important contributions to the UNL development are:

- a converter for English and Russian supporting latest specifications and proposals,
- common dictionary of UNL concepts, which is already used by the Russian and French groups and includes UNL↔French and UNL↔Russian dictionaries,
- an improved system of UNL attributes to encode modality and evidentiality.

3.1 Non-tree structures

There are several cases when the enconverter can create non-tree structures. 1) The UNL relations *and*, *or* and argument relations of participles are reversed, i.e. go from syntactic daughters to their parent nodes. In some sentences, as in Fig. 6, this results in non-tree graphs.

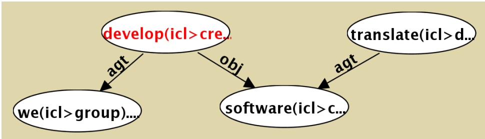

Figure 6: Non-tree graph of the phrase "We develop translating software."

2) The semantic scopes (Boguslavskij, 1996) of modifiers syntactically attached to one member of a coordinated chain but covering other members as well can be represented either by using a hypernode or several UNL links connecting the modifier with each of the modified chain members. For example, the scope of modifier "*fresh*" in (2) "*The RKI has continued to warn Germans to avoid eating fresh cucumbers, tomatoes and lettuce.*" includes all three members of the chain "*cucumbers, tomatoes and lettuce*" (See Fig.7).

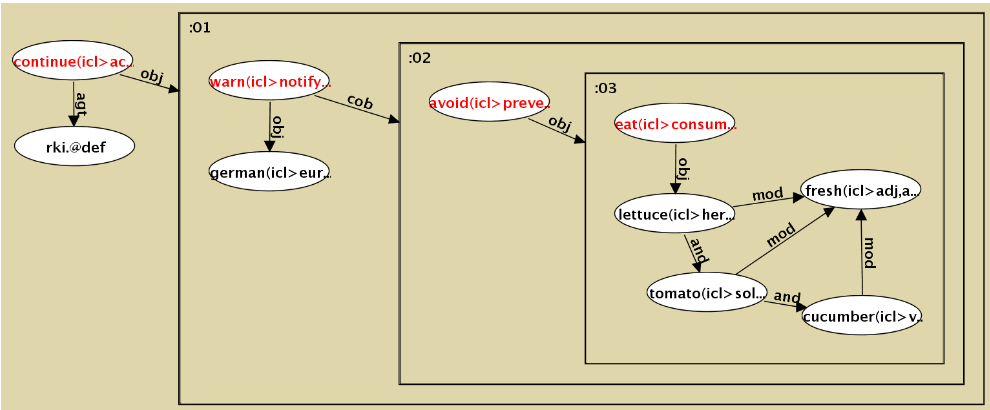

Figure 7: Non-tree graph of the phrase "The RKI has continued to warn Germans to avoid eating fresh cucumbers, (fresh) tomatoes and (fresh) lettuce."

3) Referential relations[3] by their nature disregard the tree structure. An anaphoric word and its antecedent or two different namings of the same object are always placed in different subtrees. Connecting them in means creation of a non-tree graph. For example, the sentence (3) "*The cat eats the mouse the cat catches*" discussed by Etienne Blanc (Blanc, 2005) contains two mentions of the same cat and should be represented in UNL by a circle graph, where both "*cat*" and "*mouse*" receive two parent predicates "*eat*" and "*catch*" (See Fig.8) while (4) "*Tom and his toy*" contains the proposed optional *ref* link between "*he*" and "*Tom*".

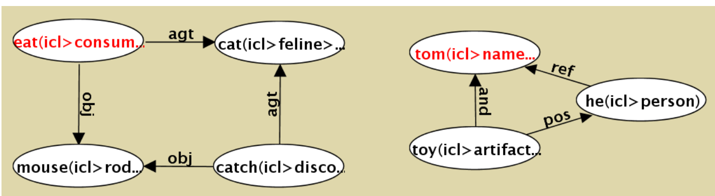

Figure 8: Non-tree graphs 3 and 4 are caused by coreference.

[3] We argue that using numeric indices to show coreference in UNL is not enough and we must introduce an extra UNL relation "ref" (referential) to support anaphoras.

3.2 Hypernodes

In UNL hypernodes are also called “scopes”. They have three different uses: 1) to represent the semantic scopes of predicates, which are broader than one word, 2) to isolate fragments of text, such as quotations, remarks, etc., inside quotes, brackets and double dashes and carry the respective attributes of formatting, 3) to segment a complex sentence into smaller parts containing individual propositions with a single predicate. Such segmentation of sentences is characteristic of our enconverter. It helps the user to view and understand large complex UNL graphs in their visual form. More about it is written in (Dikonov, 2008). The hypernodes in figures 5, 6, 9 well demonstrate the rule of segmentation and treatment of double punctuation marks (node :01 in figure 5). Hypernodes constitute a major addition to the ETAP-3 formalism which extends previous limits on the kind of supported structures.

3.3 Modality in UNL

The new enconverter supports an improved system of UNL modal attributes. It is described in detail in a separate paper (Dikonov, 2009)[4]. Its development was stimulated by the drawbacks of the old set of attributes prescribed by the (UNL specifications, 2005). The old attributes remained ambiguous and were too closely tied to the English modal verbs, which made encoding of some modal utterances in other languages inadequate. Besides, they lacked mutual organization that would enable proper translation of modals between languages having not fully compatible systems of modality. For example, some languages enforce the distinction between being physically capable of doing something and knowing how to do it. They require different modal words where English uses the same modal verb “can”.

Figure 9 shows the result of automatic parsing of the English sentence (5) “*Could you be so kind as to advise me of what could have happened?*” which demonstrates two meanings of “*could*” and the ability of the enconverter to disambiguate modal expressions and detect formulas of politeness.

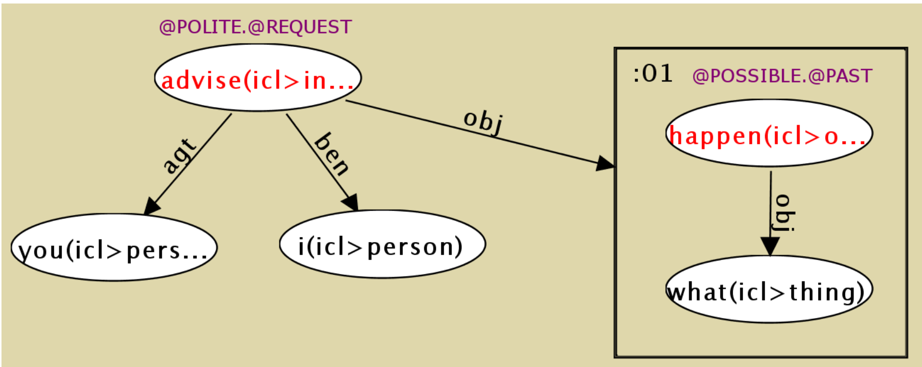

Figure 9: Disambiguation of modal verbs.

The new system of modal attributes is backwards compatible with the old one. Each UNL deconverter supporting the new attributes will be able process older graphs without any modifications because all old attributes are mapped to new ones.

---

[4] An extended version in English is currently available on request.

Current status and problems

Our UNL enconverter is able to produce semantic representation of English and Russian text but there are still things to improve. We are not able to give precision and recall scores against a large manually checked UNL corpus because such corpus has yet to be prepared. There are two problems that cause errors: wrong syntactic structures received as input and shortage of semantic knowledge in the system. The underlying syntactic parser of ETAP considers internally many syntactic trees permitted by the grammar and then selects one with the highest probability to be the correct one. However, if this final choice goes wrong, the enconverter is stuck with garbage input, unless the user switches to interactive mode and manually corrects the syntactic tree. The system has no means to judge, which of the alternative syntactic trees is the best from the semantic point of view. The other problem of insufficient semantic knowledge affects word sense disambiguation and limits the system's ability to handle complex phenomena, such as light verbs and scopes of modifiers. Only 3.1% of polysemic English words in the dictionary have word sense disambiguation rules. In order to achieve best possible performance the density of such rules must be increased. Some rules cannot be adequately formulated without extra ontological knowledge. The ontology associated with the UNL dictionary is going to provide the necessary information but it is still a work in progress.

Bibliography

Blanc E. 2005. About and around the French enconverter and the French deconverter. *Universal Networking Language: advances in theory and applications* Mexico City: National Polytechnic Institute;157 – 166

Boguslavskij, I.M. 1996. Sfera dejstvija leksicheskih edinic. *Shkola «Jazyki russkoj kul'tury», M.*

Boguslavsky I.M., Dikonov V.G. 2008. Universal Dictionary of Concepts *Proceedings of the first MONDILEX workshop "Lexicographic Tools and Techniques"*. M.; 31 – 42

Dikonov V.G. 2008. UNL Graph Structure *Informacionnye processy* Vol.8 Issue.1; 84 – 97

Dikonov V.G. 2009. Atributy modalnosti v UNL *Informacionnye tehnologii i sistemy* Bekasovo; 230 – 238

Ju.D. Apresjan, I.M. Boguslavskij, L.L. Iomdin, A.V. Lazurskij, L.G. Mitjushin, V.Z. Sannikov, L.L. Cinman, 1992 *Lingvisticheskij processor dlja slozhnyh informacionnyh sistem.* M.: Nauka.

Mel'čuk, I. A. 1974. Ob odnoj lingvisticeskoj modeli tipa "Smysl<=>Tekst" *Izvestija Akademii nauk SSSR. Serija literatury i jazyka*. M. AN SSSR, 1974. — Vol.33. Issue.5; 436 – 447.

UNL Specifications 2005. http://www.undl.org/unlsys/unl/unl2005