Creating a Universal Networking Language Module within an Advanced NLP System[1]

Igor BOGUSLAVSKY, Nadezhda FRID, Leonid IOMDIN, Leonid KREIDLIN, Irina SAGALOVA, Victor SIZOV
Computational Linguistics Laboratory
Institute for Information Transmission Problems of the Russian Academy of Sciences
19 Bol'shoj Karetnyj, 101447 Moscow, Russia
bogus(nadya,iomdin,lenya,sagalova,sizov)@iitp.ru

Abstract

A multifunctional NLP environment, ETAP-3, is presented. The environment has several NLP applications, including a machine translation system, a natural language interface to SQL type databases, synonymous paraphrasing of sentences, syntactic error correction module, and a computer-assisted language learning tool. Emphasis is laid on a new module of the processor responsible for the interface with the Universal Networking Language, a recent product by the UN University intended for the facilitation of multilanguage, multiethnic access to communication networks such as WWW. The UNL module of ETAP-3 naturally combines the two major approaches accepted in machine translation: the transfer-based approach and the interlingua approach.

1. Introductory Remarks

ETAP-3 is a multipurpose NLP environment that was conceived in the 1980s and has been worked out in the Institute for Information Transmission Problems, Russian Academy of Sciences (Apresjan *et al*. 1992, Boguslavsky 1995). The theoretical foundation of ETAP-3 is the Meaning ⇔ Text linguistic theory by Igor' Mel'čuk and the Integral Theory of Language by Jurij Apresjan.

ETAP-3 is a non-commercial environment primarily oriented at linguistic research rather than creating a marketable software product. The main focus of the research carried out with ETAP-3 is computational modelling of natural languages. This attitude explains our effort to develop the models in a way as linguistically sound as possible. We strive to incorporate into the system much linguistic knowledge irrespective of whether this knowledge is essential for better text processing (e.g. machine translation) or not. In particular, we want our parser to produce what we consider a correct syntactic representation of the sentence – first of all because we believe that this interpretation is a true fact about the natural language. We have had many occasions to see that in the long run the theoretical soundness and completeness of linguistic knowledge incorporated in an NLP application will pay.

All NLP applications in ETAP-3 are largely based on an original system of three-value logic and use an original formal language of linguistic descriptions, FORET.

2. ETAP-3: Modules, Features, Design, Implementation

2.1 ETAP-3 Modules

The major NLP modules of ETAP-3 are as follows:

- High Quality Machine Translation System
- Natural Language Interface to SQL Type Databases
- System of Synonymous Paraphrasing of Sentences
- Syntactic Error Correction Tool
- Computer-Aided Language Learning Tool
- Tree Bank Workbench

Another module, a new UNL converter responsible for the interface with the Universal Networking Language, a recent product designed

[1] The research reported here was in part supported by a grant (No 99-06-80277) from the Russian Foundation for Fundamental Research, whose assistance is gratefully acknowledged.

by the UN University, is discussed in detail in Section 3.

2.1.1. ETAP-3 MT System

The most important module of ETAP-3 is the MT system that serves five language pairs: (1) English-Russian, (2) Russian – English, (3) Russian – Korean. (4) Russian – French, and (5) Russian – German.

By far the most advanced are the first two of these pairs. The system disposes of 50,000-strong so-called combinatorial dictionaries of Russian and English that contain syntactic, derivational, semantic, subcategorization, and collocational information. The system relies on comprehensive grammars of the two languages.

For the other language pairs smaller scale prototypes are available.

ETAP-3 is able to present multiple translations when it encounters an ambiguity it cannot resolve. By default, the system produces one parse and one translation that it considers the most probable. If the user opts for multiple translation, the system remembers the unresolved ambiguities and provides all mutually compatible parses and lexical choices. To give one example from the real output: the sentence *They made a general remark that...* , when submitted to the multiple translation option, yielded two Russian translations that correspond to radically different syntactic structures and lexical interpretations: (a) *Oni sdelali obshchee zamechanie, chto...* (≈ They made some common remark that ...) and (b) *Oni vynudili generala otmetit', chto...* (≈ They forced some general to remark that ...).

2.1.2. Natural Language Interface to SQL Type Databases

This ETAP-3 module translates freely worded human queries to a database from Russian or English into SQL expressions. It can also produce the reverse generation of a NL query from an SQL expression.

2.1.3. System of Synonymous Paraphrasing

The module is designed for linguistic experiments in obtaining multiple meaning-retaining paraphrases of Russian and English sentences. The paraphrasing is based on the concept of lexical functions, one of the important innovations of the Meaning ⇔ Text theory. The following example shows the kind of paraphrases that can be produced by the module:

(1) *The director ordered John to write a report – The director gave John an order to write a report – John was ordered by the director to write a report – John received an order form the director to write a report.*

It is a very promising direction of linguistic research and development that can be applied in a wide range of activities, including language learning and acquisition, authoring, and text planning. Besides that, lexical functions are used for ensuring adequate lexical choice in machine translation and in the UNL module.

2.1.4. Syntactic Error Correction Tool

The module operates with Russian texts in which it finds a wide range of errors in grammatical agreement as well as case subcategorization and offers the user the correct version.

2.1.5. Computer-Aided Language Learning Tool

The module is a standalone software application constructed as a dialogue type computer game intended for advanced students of Russian, English, and German as foreign languages who wish to enrich their vocabulary, especially to master the collocations of these natural languages and their periphrastic abilities. The tool relies on the apparatus of lexical functions. It can also be used native speakers of the three languages interested in increasing their command of the vocabulary (such as journalists, school teachers, or politicians).

2.1.6. Tree Bank Workbench

This is the module that utilizes the ETAP-3 dictionaries, its morphological analyzer and the parser to produce a first-ever syntactically tagged corpus of Russian texts. It is a mixed type application that combines automatic parsing with human post-editing of tree structure.

2.2. Major Features

The following are the most important features of the whole ETAP-3 environment and its modules:

- Rule-Based Approach
- Stratificational Approach
- Transfer Approach
- Syntactic Dependencies
- Lexicalistic Approach
- Multiple Translation
- Maximum Reusabilty of Linguistic Resources

In the current version of ETAP-3, its modules that process NL sentences are strictly rule-based. However, in a series of recent experiments, the MT module was supplemented by an example-based component of a translation memory type and a statistical component that provides semiautomatic extraction of translation equivalents from bilingual text corpora (see Iomdin & Streiter 1999).

ETAP-3 shares its stratificational feature with many other NLP systems. It is at the level of the normalized, or deep syntactic, structure that the transfer from the source to the target language takes place in MT.

ETAP-3 makes use of syntactic dependency trees for sentence structure representation instead of constituent, or phrase, structure.

The ETAP-3 system takes a lexicalistic stand in the sense that lexical data are considered as important as grammar information. A dictionary entry contains, in addition to the lemma name, information on syntactic and semantic features of the word, its subcategorization frame, a default translation, and rules of various types, and values of lexical functions for which the lemma is the keyword. The word's **syntactic features** characterize its ability/non-ability to participate in specific syntactic constructions. A word can have several syntactic features selected from a total of more than 200 items. **Semantic features** are needed to check the semantic agreement between the words in a sentence. The **subcategorization frame** shows the surface marking of the word's arguments (in terms of case, prepositions, conjunctions, etc.). **Rules** are an essential part of the dictionary entry. All the rules operating in ETAP-3 are distributed between the grammar and the dictionary. Grammar rules are more general and apply to large classes of words, whereas the rules listed or simply referred to in the dictionary are restricted in their scope and only apply to small classes of words or even individual words. This organization of the rules ensures the self-tuning of the system to the processing of each particular sentence. In processing a sentence, only those dictionary rules are activated that are explicitly referred to in the dictionary entries of the words making up the sentence. A sample dictionary entry fragment for the English noun *chance* illustrates what was said above:

[1] *CHANCE1*
[2] POR:S
[3] SYNT:COUNT,PREDTO,PREDTHAT
[4] DES:'FACT','ABSTRACT'
[5] D1.1:*OF*,'PERSON'
[6] D2.1:*OF*,'FACT'
[7] D2.2:*TO2*
[8] D2.3:*THAT1*
[9] SYN1:*OPPORTUNITY*
[10] MAGN:*GOOD1/FAIR1/EXCELLENT*
[11] ANTIMAGN:*SLIGHT/SLIM/POOR/LITTLE1/ SMALL*
[12] OPER1:*HAVE/STAND1*
[13] REAL1-M:*TAKE*
[14] ANTIREAL1-M:*MISS1*
[15] INCEPOPER1:*GET*
[16] FINOPER1:*LOSE*
[17] CAUSFUNC1:*GIVE<TO1>/GIVE*
[18] ZONE:R
[19] TRANS:*SHANS/SLUCHAJ*
[20] REG:TRADUCT2.00
[21] TAKE:X
[22] LOC:R
[23] R:COMPOS/MODIF/POSSES
[24] CHECK
[25] 1.1 DEP-LEXA(X,Z,PREPOS,BY1)
[26] N:01
[27] CHECK
[28] 1.1 DOM(X,*,R)
[29] DO
[30] 1 ZAMRUZ:Z(PO1)
[31] 2 ZAMRUZ:X(SLUCHAJNOST')
[32] N:02
[33] CHECK
[34] 2.1 DOM(X,*,*)
[35] DO
[36] 1 ZAMRUZ:Z(SLUCHAJNO)
[37] 2 STERUZ:X
[38] TRAF:RA-EXPANS.16
[39] LA:THAT1
[40] TRAF:RA-EXPANS.22

Line [2] - part of speech: a noun.
Line [3] - the list of syntactic features.
Line [4] - the list of semantic features.
Lines [5] - [8] - the subcategorization frame.
Lines [9] - [17] - the list of lexical functions used to describe restricted lexical co-occurrence.
Line [18] – marks the end of the application-independent information and beginning of the information used in the English-Russian translation.
Line [19] – default translation into Russian.
Lines [20] – [37] – a rule for translating the phrase *by chance* in different contexts.
Lines [38] – [39] – a reference to the rule which introduces a semantically empty conjunction (*that: a chance that we obtain a grant*) .
Line [40] – a reference to the rule which introduces particle *to* (*a chance to win*).

2.3. General Architecture of the ETAP-3 environment.

To give a general idea of how the ETAP-3 NLP operates, we show here the layout of the MT module (Fig. 1). In a way, all the other modules can be viewed as this module's derivatives.

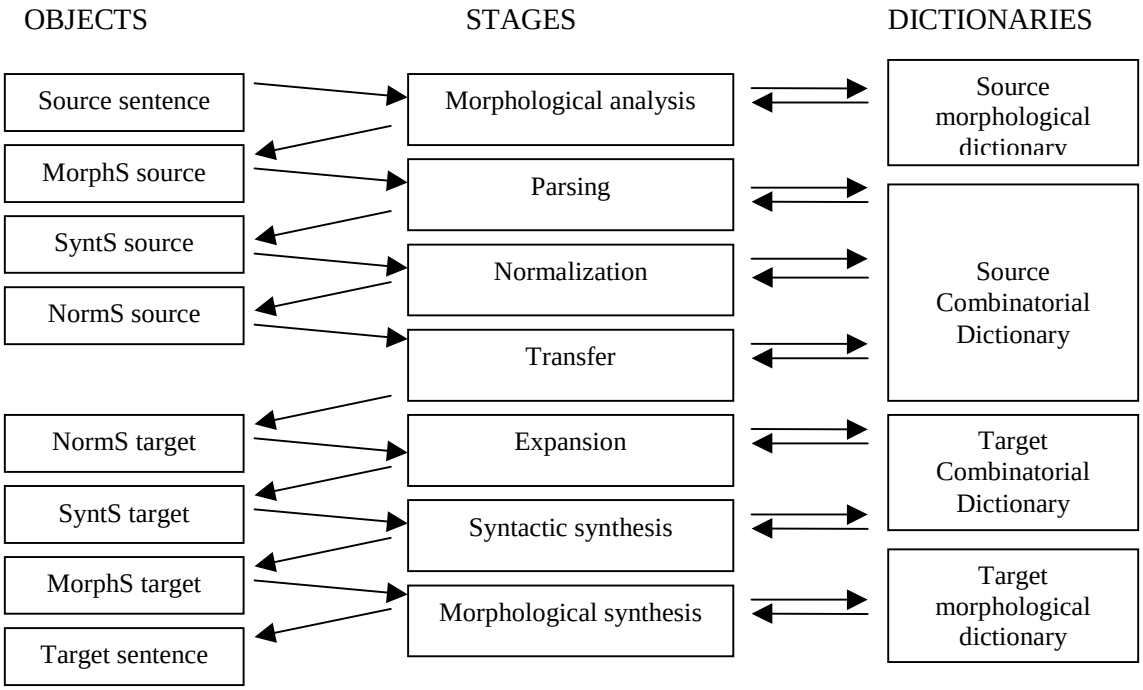

Fig.1

2.4. Implementation

The ETAP-3 environment has been implemented on a PC under Windows NT 4.0 environment. The environment has a number of auxiliary tools, including a sophisticated lexicographer's toolkit that allows the developers and the users to effectively maintain and update the ETAP-3 dictionaries.

3. The UNL Interface

3.1 Aims and scenario

The UNL project has a very ambitious goal: to break down or at least to drastically lower the language barrier for the Internet users. With time and space limitations already overcome, the Internet community is still separated by language boundaries. Theoretically, this seems to be the only major obstacle standing in the way of international and interpersonal communication in the information society. This is why the problem of the language barrier on the Internet is perceived as one of the global problems of mankind, and a project aiming to solve this problem has been initiated under the UN auspices – by the Institute of Advanced Studies of the United Nations University.

Started in 1996, the project currently embraces 15 universities and research institutions from Brazil, China, Egypt, France, Germany, India, Indonesia, Italy, Japan, Jordan, Latvia, Mongolia, Russia, Spain, and Thailand. In the following years more groups are expected to join, so that in the long run all languages of the UN member states will be covered.

The idea of the project is as follows. An interlingua has been developed which has sufficient expressive power to represent relevant information conveyed by natural languages. This interlingua entitled Universal Networking Language (UNL) has been proposed by H. Uchida (UNU/IAS). For each natural language, two systems should be developed: a "deconverter" capable of translating texts from UNL to this NL, and an "enconverter" which has to convert NL texts

into UNL. It should be emphasized that the procedure of producing a UNL text is not supposed to be fully automatic. It will be an interactive process with the labor divided between the computer and a human expert ("writer") in UNL.

This paradigm makes UNL radically different from conventional machine translation. Due to the interactive enconversion, the UNL expression, which serves as input for generation, can be made as good as one wishes. The UNL writer will edit the rough result proposed by the enconverter, correct its errors, eliminate the remaining ambiguities. He/she can run a deconverter of his own language to test the validity of the UNL expression obtained and then refine it again till one is fully satisfied with the final result.

Another important distinction from MT systems is that the interlingua representation of texts will be created and stored irrespectively of its generation into particular languages. UNL can be seen as an independent means of meaning representation. UNL documents can be processed by indexing, retrieval and knowledge extraction tools without being converted to natural languages. Generation will only be needed when the document has reached the human user.

A deconverter and an enconverter for each language form a Language Server residing in the Internet. All language servers will be connected in the UNL network. They will allow any Internet user to deconvert a UNL document found on the web into his/her native language, as well as to produce UNL representations of the texts he/she wishes to make available to multiethnic public.

3.2 UNL language

We cannot describe the UNL language here in all details: this topic deserves a special paper which will hopefully be written by the author of the language design – Dr. Hiroshi Uchida. We will only characterize it to the extent necessary for the description of our deconversion module. Full specification of UNL can be found at *http://www.unl.ias.unu.edu/*.

UNL is a computer language intended to represent information in a way that allows to generate a text expressing this information in a very large number of natural languages. A UNL expression is an oriented hyper-graph that corresponds to a NL sentence in the amount of information conveyed. The arcs of the graph are interpreted as semantic relations of the type *agent, object, time, place, instrument, manner,* etc. The nodes of the graph are special units, the so-called Universal Words (UW) interpreted as concepts, or groups of UWs. The nodes can be supplied with attributes which provide additional information on their use in the given sentence, e.g. ***@imperative, @generic, @future, @obligation***.

Each UW is represented as an English word that can be optionally supplied with semantic specifications to restrict its meaning. In most cases, these specifications locate the concept in the knowledge base. It is done in the following way: UW *A(icl>B)* is interpreted as 'A is subsumed under the category B'. For example, the UW *coach* used without any restrictions denotes anything the English *coach* can denote. If one wants to be more precise, one can use restrictions: *coach(icl>transport)* denotes a bus, *coach (icl>human)* denotes a trainer and *coach (icl>do)* denotes the action of training. In a sense, the apparatus of restrictions allows to represent UWs as disambiguated English words. On the other hand, restrictions allow to denote concepts which are absent in English. For example, in Russian there is a large group of motion words, whose meaning incorporates the idea of the mode of locomotion or transportation: *priletet'* 'come by flying', *priplyt'* 'come by ship', *pripolzti* 'come by crawling', *pribezhat'* 'come running', etc. English has no neutral words to denote these concepts. Still, on the basis of English one can construct UWs that approximate required concepts, e.g. *come(met>ship)* is interpreted as 'come and the method of coming is a ship'.

Here is an example of a UNL expression for the sentence

(2) *However, language differences are a barrier to the smooth flow of information in our society.*

Each line is an expression of the kind *relation(UW1, UW2).* For simplicity, UWs are not supplied with restrictions.

aoj(barrier.@entry.@present.@indef.@however, difference.@pl)
mod(barrier.@entry.@present.@indef.@however, flow.@def)
mod(difference.@pl, language)
aoj(smooth, flow.@def)
mod(flow.@def, information)
scn(flow.@def, society)
pos(society, we)

Relations used: **aoj** - a relation that holds between a thing and its state, **mod** - a relation

between a thing and its modifier, **scn** - a relation between an event or a state and its abstract location, **pos** - a relation between a thing and its possessor. Attributes: **@entry** - denotes the top node of the structure, **@present** - present tense, **@def** - definite NP, **@pl** - plural, **@however** - a modal meaning corresponding to English *however*.

3.3. UNL – Russian deconversion by means of ETAP-3

As was shown in Section 1, ETAP-3 is a transfer-based system where the transfer is carried out at the level of the Normalized Syntactic Structure (NormSS). This level is best suited for establishing correspondence with UNL, as UNL expressions and NormSS show striking similarities. The most important of them are as follows:

1. Both UNL expressions and NormSSs occupy an intermediate position between the surface and the semantic levels of representation. They roughly correspond to the so-called deep-syntactic level. At this level the meaning of the lexical items is not decomposed into the primitives, and the relations between the lexical items are language independent;
2. The nodes of both UNL expressions and NormSSs are terminal elements (lexical items) and not syntactic categories;
3. The nodes carry additional characteristics (attributes);
4. The arcs of both structures are non-symmetrical dependencies.

At the same time, UNL expressions and NormSSs differ in several important respects:

1. All the nodes of NormSSs are lexical items, while a node of a UNL expression can be a sub-graph;
2. Nodes of a NormSS always correspond to one word sense, while UWs may either be broader or narrower than the corresponding English words:

2.1. they can cover a meaning area that corresponds to several different word senses at a time (see above);

2.2. they can correspond to a free word combination (e.g. *computer-based* or *high-quality*);

2.3. they can correspond to a word form (e.g. *best* which a form of *good* or *well*);

2.4. they can denote a concept that has no direct correspondence in English (see above).

3. A NormSS is the simplest of all connected graphs - a tree, while a UNL expression is a hyper-graph. Its arcs may form a loop and connect sub-graphs;
4. The relations between the nodes in a NormSS are purely syntactic and are not supposed to convey a meaning of their own, while the UNL relations denote semantic roles;
5. Attributes of a NormSS mostly correspond to grammatical elements, while UNL attributes often convey a meaning that is expressed both in English and in Russian by means of lexical items (e.g. modals);
6. A NormSS contains information on the word order, while a UNL expression does not say anything to this effect.

The NormSS of the sentence (2) looks as follows:

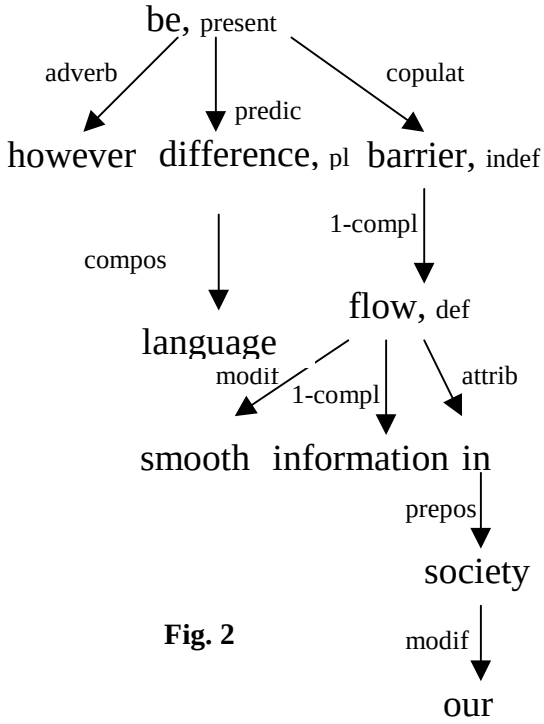

Fig. 2

As UNL makes use of English lexical labels, it is expedient to bridge the gap between UNL and Russian via English NormSS which actually serves as an Intermediate Representation (IR). In this case the UNL - Russian interface will be the simplest. After the English NormSS has been reached, conventional ETAP English-to-Russian machine translation mode of operation can be used.

The UNL-to-Russian module carries out the following three steps:

1. Transfer from UNL to the intermediate representation (IR).

2. Transfer from the IR to the Russian normalized syntactic structure (NormSS-R).
3. Generation of a Russian sentence from the NormSS-R.

The architecture of the UNL-Russian deconverter is shown in Fig. 3.

It follows from the previous discussion that the UNL - NormSS interface should solve the following five tasks:

1. An appropriate English lexeme for every UW should be selected where it is possible; a Russian lexeme will be provided by the ETAP English - Russian transfer dictionary. If no appropriate English word can be found for a UW, other means of expression should be found.
2. UNL syntactic relations should be translated, either by means of ETAP relations or with the help of lexical items.
3. UNL attributes should be translated, either by means of grammatical features or with the help of lexical items (e.g. @however - *however*).
4. UNL graph should be converted in a tree.
5. Word order should be established.

The first and (partly) the second tasks are solved by means of the information stored in the UW - English and English combinatorial dictionaries. All the rest (tasks 2 to 5) is done by the rules written in the logical-based FORET formalism.

Let us give one example to illustrate the transformation of UNL relations into NL words. UNL has a **tim** relation that holds between an event and its time. As is known, the choice of appropriate words to express this relation is to a large extent determined by lexical properties of the word denoting time; cf. *on Monday, at midnight, in summer, during the war,* etc. In ETAP-3 all these cases are treated as the lexical function LOC denoting (temporal) locality (on lexical functions see 2.1.3). The values of all lexical functions are given in the lexicon in the entries of their arguments (see an example in 2.2 above). While processing the UNL expression, the tim relation is linked to the lexical function LOC which allows to find a correct preposition, both in English and in Russian.

3.4. Current state and prospects for the future

The module of Russian deconversion is operational and can be tested at

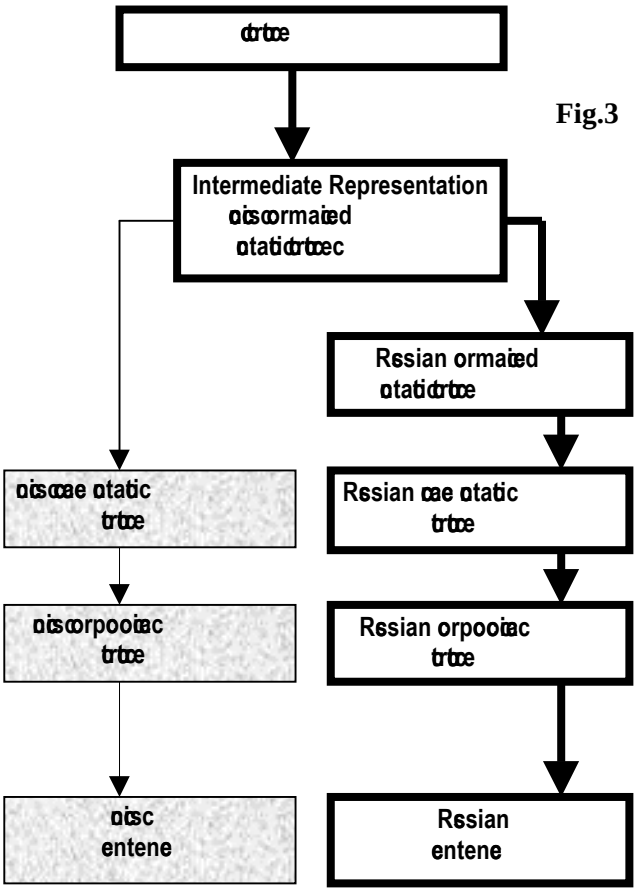

Fig.3

http://proling.iitp.ru/Deco. We plan to put it to general use by autumn 2000. The interactive enconversion module will be our next concern.

As shown in Fig. 3, the interface between UNL and Russian is established at the level of the English NormSS. At this point ETAP English-to-Russian machine translation facility can be switched which carries through the phases of transfer and Russian generation. This architecture allows to obtain English generation for relatively cheap, as ETAP has a Russian-to-English mode of operation as well. First experiments in this direction have been carried out which proved quite promising.

References

Apresjan Ju.D., I.M.Boguslavsky, L.L.Iomdin *et al*. (1992). ETAP-2: The Linguistics of a Machine Translation System. // META, Vol. 37, No 1, pp. 97-112.

Boguslavsky I.(1995). A bi-directional Russian-to-English machine translation system (ETAP-3). // Proceedings of the Machine Translation Summit V. Luxembourg.

Iomdin L.& O. Streiter. (1999). Learning from Parallel Corpora: Experiments in Machine Translation. // Dialogue'99: Computational Linguistics and its Applications International Workshop. Tarusa, Russia, June 1999. Vol.2, pp. 79-88.