ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the Meaning ⇔Text Theory[1].

Jurij D. Apresian, Igor M. Boguslavsky, Leonid L. Iomdin, Alexander V. Lazursky, Vladimir Z. Sannikov, Victor G. Sizov, Leonid L. Tsinman

Laboratory of Computational Linguistics
Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow
{apr,bogus,iomdin,lazur,san,sizov,cin}@cl.iitp.ru

Résumé

On présente la description du processeur linguistique ETAP-3, qui constitue un environnement multifonctionnel pour le traitement du langage naturel. Cet environnement, essentiellement basé sur la théorie Sens ⇔ Texte, offre plusieurs applications pratiques, dont un système de traduction automatique, un système de paraphrase synonymique, un système d'annotation syntaxiques d'un corpus de textes, un interface UNL, un système d'enseignement des langues étrangères assisté par ordinateur, un interface permettant de dialoguer en langage naturel avec les bases de données de type SQL et un système de correction automatique des erreurs syntaxiques dans le texte. L'article, qui contient un bref aperçu de toutes les applications mentionnées, porte un accent particulier sur la traduction automatique qui est de loin l'application la plus développée de cet environnement.

Abstract

A multifunctional NLP environment, ETAP-3 linguistic processor, is presented. The environment, largely based on the Meaning ⇔ Text Theory, offers several NLP applications, including a machine translation system, a module of synonymous paraphrasing of sentences, a tagger for syntactic annotation of text corpora, a Universal Networking Language interface, a computer-assisted language learning tool, a natural language interface to SQL type databases, and a syntactic error correction module. While all applications are briefly discussed, emphasis is laid on machine translation, as it is by far the most advanced application of all.

[1] The authors are grateful to the Russian Fund of Fundamental Research (grants No 01-07-90405 and 02-06-80085), who partially supported this work.

Keywords – Mots Clés

traitement du langage naturel, traduction automatique, analyse et synthèse morphologique, analyse syntaxique, dépendance syntaxique, dictionnaire combinatoire, fonctions lexicales

Natural Language Processing, Machine Translation, Morphological Analysis and Synthesis, Parsing, Syntactic Dependency, Combinatorial Dictionary, Lexical Functions

1 Introductory Remarks

The multifunctional ETAP-3 linguistic processor, developed by the Computational Linguistics Laboratory (CLL) in Moscow (see e.g. Apresjan *et al.* 1992a,b, 1993), is the product of more than two decades of laboratory research and development in the field of language modeling. The most important features of the processor are as follows.

(1) ETAP-3 is based on the **general linguistic framework** of the Meaning ⇔ Text theory, proposed by Igor Mel'čuk [e.g. Mel'čuk 1974]; a natural complement to this theory was furnished by the concept of systematic lexicography and integrated description of language proposed by Jurij Apresjan [Apresjan 1995, 2000], who was head of the CLL for 20 years. As MTT strives to describe the most fundamental linguistic abilities of man – those of producing natural language texts and understanding them, it is well suited for the development of all major types of NLP applications. Surprisingly enough, only a few large-scale efforts to implement MTT in linguistic engineering tasks have been made so far, and as far as MT is concerned, ETAP-3 seems to be the only endeavour in this direction of activity.

(2) ETAP-3 has **a declarative organization of linguistic knowledge**. It means that linguistic data (the grammar and the dictionary) are conceptually separated from the software that is used to process them. Due to that, linguistic knowledge is not dispersed in the software code and is therefore transparent, easy to understand and maintain.

(3) One of the major components of ETAP-3 is **the innovative combinatorial dictionary**, ideologically based on the notion of the MTT's explanatory combinatorial dictionary. Apart from syntactic and semantic features and subcategorization frames, the dictionary entry may have rules of 8 types. Many dictionary entries contain **lexical functions** (LF).

(4) ETAP-3 makes use of a formalism based on first order predicate logic, in which all linguistic data are presented. It is rich enough to enable the formulation of highly sophisticated rules. The formalism is simple to learn and use and, with modern computers, it sufficiently fast computation.

(5) The ETAP-3 processor has **a modular architecture**. All stages of processing and all types of linguistic data are organized into modules, which warrants their reusability in many NLP applications both within and beyond ETAP-3 environment. For example, the same morphological analyzers, syntactic parsers and dictionaries are used in a variety of ETAP-3 applications. On the other hand, these modules can be (and some of them are) used in search engines, advanced information retrieval tasks, data mining, summarization, question-answering systems, and other NLP tasks. The modular architecture also facilitates a combination of the ETAP-3 rule-based MT module with translation memories and other statistics-based translation tools (see e.g. Carl *et al.* 2000).

At the moment, the ETAP-3 environment comprises the following main options: 1) a rule-based machine translation system; 2) a system of synonymous paraphrasing of sentences; 3) a workbench for syntactic annotation of text corpora; 4) a Universal Networking Language translation engine; 5) a natural language interface to SQL-type databases, and 6) a grammar checker. We will first discuss the machine translation system, as it is by far the most advanced application of all, and then briefly discuss other ETAP-3 options.

2 ETAP-3 Machine Translation System

2.1 Major Options

The current ETAP-3 MT options include the Russian-to-English and the English-to-Russian translation pair; and a number of prototypes: Russian – French, Russian – German, Russian – Korean and Russian – Spanish. We will focus on the Russian/English option as the remaining ones are only small experimental systems of varied advancement degrees.

2.2 General Architecture of Translation Process

The general design of the multistage MT process is presented in Fig. 1.

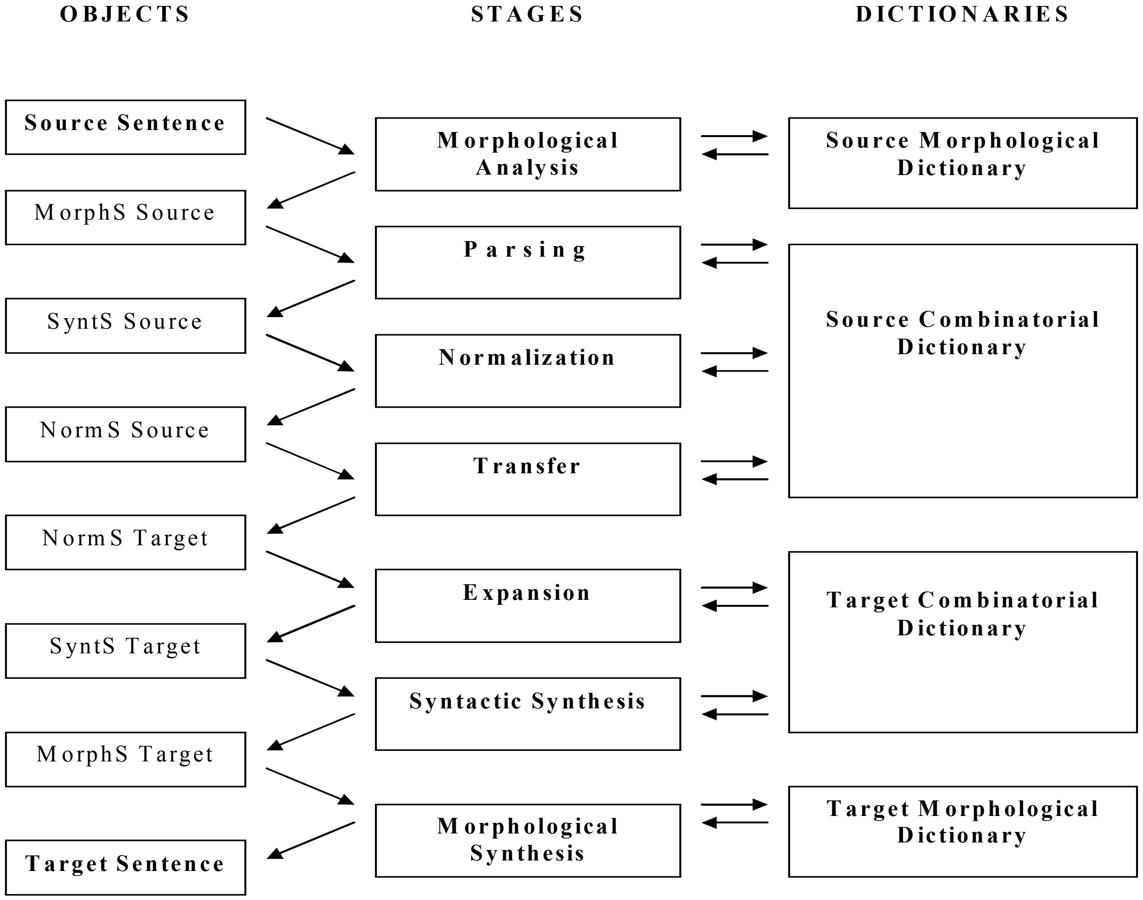

Figure 1 : General Scheme of ETAP-3 Machine Translation

In ETAP-3, translation is performed sentence by sentence. Every sentence in the source language is first morphologically analyzed, which means that every word is assigned a deep morphological representation, i.e. the lemma furnished with inflectional characteristics. If a word is morphologically and/or lexically ambiguous, it is assigned a set of morphological representations; for example, the English word *coaches* is represented as {COACH1, V, prs, 3-p, sg / COACH2, S, pl 'buses' / COACH3, S, pl 'trainers'}. Morphological analysis does not take into account any word context, so no lexical or morphological ambiguity is resolved at this stage. The sequence of all morphological representations of the words of a sentence is its morphological structure (MorphS).

ETAP-3 morphological module uses vast morphological dictionaries (the Russian dictionary counts 130,000 entries amounting to several million word forms, and the English counts 70,000 entries), and a finite-state software engine. The morphological analyzer is able to parse compound words like Russian *odinnadcatimetrovyj* 'eleven-meter' or English *bioterrorism.*

The MorphS of the source sentence is processed by a small pre-syntactic module, which partially resolves lexical and morphological ambiguity using information of close linear context. To give a simple example, if the ambiguous word *coach* is preceded by an article *the* or *a*, its verbal interpretation is excluded from further consideration, The MorphS of the source sentence, partially disambiguated by the pre-syntactic module, is then sent to the parser – the most important and sophisticated part of the system.

2.3 Parsing

The parsing module of ETAP-3 transforms the MorphS of the sentence into a classic MTT (surface) dependency tree structure[2]. The tree nodes correspond to the words of the sentence, while the arcs are labeled with names of surface syntactic relations (SSR). The parsing algorithm creates from the linear MorphS a dependency tree using syntagms, or rules that produce minimal subtrees consisting of two nodes connected with a directed arc labeled by an SSR. The set of syntagms consists of several hundred rules for each of the two working languages, written in a specially designed formalism, FORET, used for all types of ETAP rules. Normally, every syntagm describes a specific binary syntactic construction (e.g. nominal subject + verbal predicate as in *war stinks*, noun plus adjectival modifier, as in *fair play*, numeral plus noun, as in *seven seas*, etc.).

Parser operation consists of several phases. In the first phase, syntagms create for the given MorphS all possible syntactic hypotheses, or links, using only linear word order information. In most cases, the set of hypothetical links produced for a MorphS is much (by an order of magnitude) greater than the set of links needed to build a tree. Accordingly, at subsequent phases of parsing extraneous links are eliminated with the help of several powerful filtering mechanisms. The main filters include projectivity restrictions, universal and local tree constraints, non-repeatability conditions for certain SSRs, etc. An important innovation introduced to parsing theory and practice by ETAP-3 is an original mechanism of forceful

[2] The only difference between the ETAP-3 dependency tree and the standard MTT dependency tree is the fact that ETAP-3 parser **retains** in the tree the linear order of the source sentence.

detection of the top node that resorts to empirical preference rules based on close linguistic observation of syntactic structures.

If a sentence is syntactically ambiguous, the parser is able to produce several SyntS corresponding to different readings. Two screenshots in Fig. 2 below represent two SyntS for an ambiguous sentence (1) *He made a general remark that everything was OK.* The left-hand SyntS corresponds to the reading ‘He forced some general to remark that everything was OK’ and the right-hand one to the reading ‘He remarked in a general way that everything was OK’.

To optimize the parsing process, syntagms are arranged into three types: general syntagms operating on each sentence processed, template syntagms referred to in dictionary entries of restricted word classes, and dictionary syntagms located directly in the entries of syntactically salient words (auxiliaries, particles, conjunctions etc.). This type of rule arrangement is applied in all ETAP-3 phases and modules with the exception of morphological ones.

Recent innovations in the ETAP-3 parser include 1) a system of empirical weights dynamically assigned to the elements of the dependency tree at earlier stages of the parsing process (Iomdin *et al. 2002*), 2) a module of interactive resolution of lexical ambiguity that involves participation of a human expert, and 3) a module of preference rules based on statistics learned from syntactically annotated corpora (see below, Section 4).

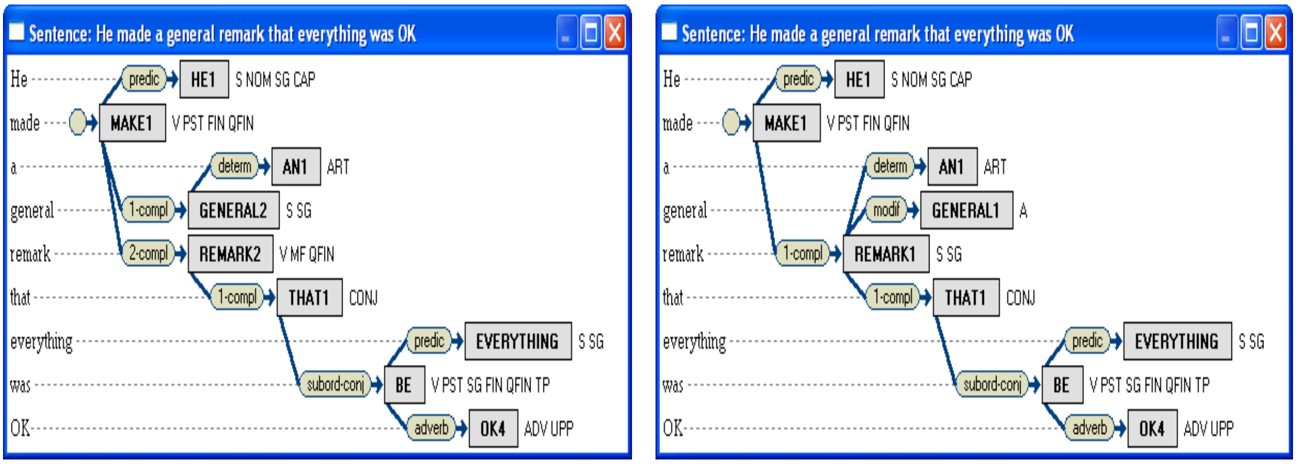

Figure 2 . SyntS for Two Readings of Sentence (1).

The ready SyntS is sent to the SyntS normalization module that is used to strip the SyntS structure of some of the specific features of the source language. Typical normalization rules merge into single nodes verbal expressions formed with auxiliaries, remove from SyntS strongly governed prepositions and conjunctions, articles (while transferring information on definiteness to the head of the nominal phrase), and identify arguments and values of LFs. Besides, these rules delete syntactically conditioned morphological features (like gender, number or case in agreed adjectives, number and person of finite verbs etc.) The output of the normalization module is called Normalized Syntactic Structure, or NormS. As a matter of fact, NormS is an approximation to the concept of MTT Deep SyntS but it does not make use of deep syntactic relations: NormS relations are inherited from SyntS.

2.4 Transfer

The transfer proper, i.e. the transition from the source language to the target language, is performed at the level of NormS. Even though the translation process does not resort to classic semantic structures, the NormS provides sufficient control of sentence semantics as many of the SSRs are semantically motivated and the nodes carry semantic data inherited from the combinatorial dictionaries of the source language. As a result of the transfer phase operation, the NormS of the source language is replaced by a NormS of the target language, in which all nodes represent the words of the target language and the arcs are labeled with target SSR names. Special provisions are made to enable due processing of LF values.

The target NormS is sent to a refinement module, called expansion, which fulfils operations inverse to the ones performed by the normalization module. In particular, it generates analytical verb forms, introduces articles and strongly governed prepositions and conjunctions, and ensures the right word order of the target sentence. The resulting expanded target SyntS is almost ready for the next-but-last phase of translation – syntactic synthesis, which produces the lacking morphological features (as required by agreement or government rules) and prepares ground for the final phase of translation – morphological generation that uses the target morphological dictionary to generate real word forms and produce the target sentence. In our example, two Russian equivalents will be produced: *On vynudil generala zamečat', čto vse bylo xorošo* and *On sdelal obščee zamečanie, čto vse bylo xorošo.*

2.5 Combinatorial Dictionaries

As already mentioned, an important characteristic feature of ETAP-3 is high reusability of its linguistic resources. Combinatorial dictionaries, which are slightly reduced (they provide no lexicographic definitions) but fully formalized versions of explanatory combinatorial dictionaries (ECD) of MTT, are the most important and valuable type of reusable ETAP-3 resources. In particular, the Russian combinatorial dictionary is used as the source dictionary in the Russian-to-English translation and as the target dictionary in the opposite direction of translation. For the English combinatorial dictionary, the reverse is true. Both dictionaries are used in several ETAP-3 options in addition to MT.

Currently, both dictionaries contain about 65,000 lexical entries each and offer rich and versatile information on syntactic and semantic features of the word, its government pattern, and, importantly, values of LFs for which the lemma is the argument. Lexical entries may contain whole parsing and transfer rules and references to such rules. The entry is divided into several zones. The first, universal, zone contains data independent of any application, while all the other zones present information referring to specific options.

2.6 Samples of Machine Translation Performed by ETAP-3

To illustrate the performance of the MT engine, we give a few uncommented and uncorrected examples of Russian-to-English translations. The material was online news by ITAR-TASS of October 26, 2001 and February 27, 2003).

(1) Угроза для рядовых американцев заразиться сибирской язвой через инфицированное письмо ничтожно мала, считают специалисты. The threat for ordinary Americans to be infected with an anthrax across an infected letter is negligibly small, specialists consider

(2) ФБР склоняется к тому, что распространение писем, содержащих споры сибирской язвы - дело рук внутренних террористов. FBI is inclined toward the fact that the dissemination of letters, containing the spores of anthrax, is an affair of inner terrorists' hands

(3) Израиль принял решение об отводе войск с палестинских территорий, начиная с субботы. Israel has made a decision on a withdrawal of troops from the Palestinian territories beginning with Saturday

(4) Президент Пакистана выступил за скорейшее завершение Западом военной операции против Афганистана. The president of Pakistan has supported the quickest termination by the West of the military operation against Afghanistan

(5) Япония объявила об отмене санкций против Индии и Пакистана, которые ранее были введены в качестве наказания за их ядерные испытания. Japan has announced a cancellation of sanctions against India and Pakistan which earlier had been introduced as a punishment for their nuclear tests.

(6) В офисе пакистанской авиакомпании в Карачи саперы обезвредили мощную бомбу. Sappers in an office of Pakistani air company in Karachi have rendered harmless a powerful bomb

(7) Японская принцесса Масако провела церемонию "одевания пояса", стремясь получить поддержку мистических сил в предстоящих родах. The Japanese princess Masako has held a ceremony "of a robing of a belt" striving to find a support of mystic forces in the forthcoming childbirth

(8) Запланированный на 8-10 ноября в Алма-Ате саммит стран- участниц Совещания по мерам доверия в Азии перенесен. The summit, planned for 8-10 November in Alma-Ata, of countries-participants of a Conference on measures of confidence in Asia is transferred

(9) Олимпийский комитет США объявит города, которые будут бороться за право принять летние Игры 2012 года. The Olympic committee of the USA will announce the cities which will struggle for right to receive summer Games of 2012

(10) ОПЕК должна сократить производство нефти на миллион баррелей для стабилизации цен на это сырье, заявил министр нефти Катара. OPEC must reduce the production of petroleum by a million barrels for a stabilization of the prices for this raw material, the minister of petroleum of Qatar has declared.

(11) Иракские оппозиционеры высказались против установления контроля США над Ираком после смещения Саддама Хусейна. The Iraqi oppositionists have spoken out against an establishment of a control of the USA over Iraq after a displacement of Saddam Hussein

(12) Французский "Мираж" выполнил два полета над Ираком в рамках помощи инспекторам ООН. The French "Mirage" has executed two flights over Iraq within the framework of the help to the inspectors of the UN.

3 Synonymous Paraphrasing

In mid-1990s, an experimental system of paraphrasing was developed within the ETAP-3 framework. It differed from MT in that paraphrasing was thought of as translation within the given language, in our case Russian. Consequently, the system of paraphrasing resorts to all stages mentioned in Fig. 1, except Transfer. In addition, it falls back upon three sets of rules: (1) LF interpretation of SyntS, (2) canonization rules, (c) paraphrasing rules proper.

To identify the LFs in the processed sentence, rules (1) use three types of data: (a) SyntS, (b) information on LFs assigned to words in the combinatorial dictionary, (c) the definitions of LFs. For instance, SyntS of the sentence *Filosofy podvergli aeto ponjatie tshchatel'nomu analizu* 'Philosophers subjected this concept to a thorough analysis' contains the pair of words (*podvergnut', ponjatie*) linked with SSR "1-compl" and the pair of words (*podvergnut', analiz*) linked with SSR "2-compl". The entry for *analiz* 'analysis' contains the information that the value of LF Labor1-2 is *podvergnut'* 'subject', while the definition of Labor1-2 says that it takes the keyword as its second complement. Since the SSRs in SyntS are such as required by this definition, *podvergnut'* will be identified as Labor1-2 value, and the grounds for paraphrasing will be thus prepared. Canonization rules reduce the processed sentence to its core structure, in our case, the structure of the sentence *Filosofy tshchatel'no proanalizirovali aeto ponjatie* 'Philosophers have thoroughly analyzed this concept'. Paraphrasing rules proper generate a cluster of paraphrases, with such variants as *Philosophers have made thorough analysis of this concept <notion>, This concept < notion> has undergone thorough analysis by the philosophers* etc. A series of paraphrasing experiments have produced several hundred paraphrase clusters, for the most part quite plausible.

4 Syntactic Annotation of Corpora

The module utilizes the ETAP-3 parser to produce the first syntactically tagged corpus of Russian texts. It is a mixed type application that combines automatic parsing with human post-editing of tree structures produced by the parser. To support the creation of annotated data, two tools have been designed: (1) a program for establishing sentence boundaries, called Chopper; (2) a post-editor for building, editing and managing syntactically annotated texts, called Structure Editor. As of today, the annotated Russian corpus comprises texts totaling 11,000 fully tagged sentences, or 130,000 words. Sentence annotation includes complete morphological markup at the word level, and a dependency SyntS.

5 Universal Networking Language Translation Engine

One of ETAP-3 options is translation between Russian and the Universal Networking Language (UNL), put forward by H. Uchida of United Nations University. UNL is a formal language intended to represent information in a way that allows the generation of a text expressing this information in a 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 are interpreted as semantic relations like *agent, object, time, place, manner,* etc. The nodes 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*.

The Russian-UNL module of ETAP-3 (Boguslavsky *et al.* 2000) makes part of a large network of UNL modules developed for major languages by a consortium of research institutions from Brazil, China, France, Germany, India, Indonesia, Italy, Japan, Jordan, Mongolia, Spain, and Thailand. This activity is coordinated by the UN University and UNDL Foundation. All information on the UNL network can be found at http://www.undl.org. Generation of Russian sentences from UNL can be performed at http://www.unl.ru.

6 Computer-Assisted Language Learning Tool

The tool at issue is based upon two explanatory combinatorial dictionaries (ECD), Russian and English, counting up to 3,000 entries each. The dictionaries store the following information on a lexeme, systematically and uniformly arranged throughout both dictionaries: a) its name, b) its analytical definition, c) its part of speech, d) its translation into the other language, e) its lexical functions. This information forms the basis for five interactive computer-based lexical games, in which the user may select the level of linguistic difficulty. The games are as follows: (1) guess the name of the lexeme from its analytical definition offered by the computer; (2) give the translation(s) of the lexeme, offered by the computer, into the other language; (3) supply the value(s) of the lexical function offered by the computer for the lexeme chosen by the user; (4) supply the value(s) of the lexical function chosen by the user for the lexeme offered by the computer; (5) do the same with the help of a prompt in the user's mother tongue. The games are broken into three levels of difficulty depending on the complexity of linguistic material. The program is furnished with a system of assessing the user's performance on the basis of the following criteria: the number of sessions; the number of correct answers in each session; the level of difficulty. Most of the material presented in both ECDs from this option has been transferred to the main combinatorial dictionaries operating in the remaining ETAP-3 options.

7 Natural Language Interface to SQL-Type Databases

The task of this option is to translate queries formulated in Russian to a formal query language, SQL. As shown in Fig.1, text analysis performed for the purpose of MT reaches the level of the Normalized SyntS. This level is in most cases sufficient for translation. The option of NL interface to SQL-databases requires a deeper analysis, as NL units have to be interpreted in terms of a concrete database. In the analysis phase, the system produces the normalized SyntS of the query using the same modules as those used in MT. After that, the normalized SyntS is submitted to semantic analysis that produces a semantic structure (SemS) directly translatable into SQL. The SemS is a tree labeled with semantic relations. The nodes of the tree are semantic elements that directly correspond to database units.

As the user is not expected to understand SQL, he cannot evaluate the result of translation. To help him make sure that the SQL representation obtained is adequate for the initial query, an inverse generation module has been developed. It produces a Russian sentence devoid of ambiguity. If the user finds that this sentence does not answer his information need, he can reformulate the query and submit it to the system again.

8 Syntactic Error Corrector for Russian

Whilst the description of syntax in ETAP-3 is naturally oriented at standard and grammatically correct sentences, our experiments with the parser showed that some of the conditions of syntagms are to a certain extent redundant: they allow for the processing of "slightly incorrect" linguistic material. If these conditions are weakened in a specific way, the change does not actually tell on the results of the parsing procedure (see Tsinman, Sizov 2000). This fact enabled us to create a procedure, later developed into a separate syntactic correction module, which is able to parse sentences where certain agreement and control rules are violated, and prompt the user for a correction. The core of the procedure consists in producing "conditional" syntactic links that are supported by all conditions of the syntagm except the concrete agreement condition. Any such link participates in the creation of SyntS and will be part of it unless overridden by a stronger unconditional link. Normally, the presence of conditional links in SyntS signals ungrammaticality. The operation of the module can be illustrated by sentences like **Ni odna iz arabskix stran ne osudili bombardirovki* 'None of the Arabic countries condemned the bombings' where subject-verb agreement in number is violated or **Cel' vizita sostoit v podgotovke namečennoj na maj vstreče* 'The purpose of the visit is to prepare the meeting planned for May' where the governed case of the last word *vstreča* is wrongly selected.

References

Apresjan, Ju.D. (1995) *Integral'noe opisanie jazyka i sistemnaja leksikografija [An Integrated Description of Language and Systematic lexicography.]* Moscow, Jazyki russkoj kul'tury.

Apresjan, Ju. D. (2000) *Systematic Lexicography*. Oxford University Press, XVIII p., 304 p.

Apresjan Ju.D., Boguslavskij I.M., Iomdin L.L., Lazurskij A.V., Mitjushin L.G., Sannikov, V.Z., Cinman, L.L. (1992) *Lingvisticheskij processor dlja slozhnyx informacionnyx sistem. [A linguistic processor for advanced information systems.]* Moskva, Nauka. 256 p. (In Russian).

Apresjan Ju.D., Boguslavskij I.M., Iomdin L.L., Lazurskij A.V., Sannikov V.Z. and Tsinman L.L. (1992b) The Linguistics of a Machine Translation System. *Meta, 37 (1): 97-112.*

Apresjan Ju.D., Boguslavskij I.M., Iomdin L.L., Lazurskij A.V., Sannikov V.Z. and Tsinman L.L. (1993). Système de traduction automatique {ETAP}. *La Traductique. P.Bouillon and A.Clas (eds).* Montréal, Les Presses de l'Université de Montréal.

Boguslavsky I., Iomdin L., Frid N., Kreidlin L. Sagalova I., Sizov V. (2000) Creating a Universal Networking Language Module within an Advance NLP System. Proceedings of *COLING in Europe. The 18th International Conference on Computational Linguistics.* Saarbrücken. Vol. 1, pp. 83-90.

Carl M., Pease C., Streiter O., Iomdin L.L. (2000). Towards a Dynamic Linkage of Example-Based and Rule-Based Machine Translation. *Machine Translation, 15 (3). 223-257*

Iomdin L.L, Sizov V.Z, Tsinman L.L (2002). Utilisation des poids empiriques dans l'analyse syntaxique: une application en Traduction Automatique. *META, 47. (3): 351-358*

Mel'čuk I.A. (1974). *Opyt teorii lingvisticheskix modelej klassa "Smysl - Tekst". [The theory of linguistic models of the Meaning – Text Type].* Moscow, Nauka. (In Russian).

Tsinman L.L., Sizov V.G. (2000). Lingvisticheskij processor ETAP: procedury oslablenija sintaksicheskix pravil i ix ispol'zovanie. [ETAP linguistic processor: procedures of weakening syntactic rules and their use.] *Slovo v tekste i v slovare (sbornik statej k semidesjatiletiju akademika Ju. D. Apresjana.* Moscow, Jazyki russkoj kul'tury. (In Russian).