Processing of Russian by the ETAP-3 Linguistic Processor

Igor Boguslavsky

Institute for Information Transmission Problems, Russian Academy of Sciences / Universidad Politécnica de Madrid

iboguslavsky@fi.upm.es

ETAP-3 is a multipurpose NLP environment developed in the Institute for Information Transmission Problems, Russian Academy of Sciences (Apresjan *et al*. 2003). The theoretical foundation of ETAP-3 is the Meaning ⇔ Text linguistic model 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 the creation of a marketable software product. The main focus of the research pursued with ETAP-3 is computational modelling of natural languages. The language for which the model offers the largest coverage and has been developed deepest of all is Russian, a language very rich in morphology. The model relies on a comprehensive grammar, which is used in several rule-based components that perform dependency parsing, deep-structure construction, and generation.

ETAP-3 disposes of a more than 130,000 lemmas-strong morphological dictionary that accounts for inflection, and a so-called combinatorial dictionary of nearly the same size. A combinatorial dictionary entry contains, in addition to the lemma name, information on syntactic and semantic features of the word, its subcategorization frame, a default translation (into English), rules of various types, and values of lexical functions (in the sense of Mel'čuk) 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 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. As a matter of fact, all 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 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 self-tuning of the system to the processing of each particular sentence. In processing a sentence, only those dictionary rules are activated that are present, or explicitly referred to. in the dictionary entries of the words making up the sentence.

An ongoing project carried out in cooperation with the Speech synthesis laboratory of the United Institute **o**f Information Science, National Academy of Sciences of Belarus is devoted to the development of a speech synthesis module for Russian. The project has two major challenges. The first is determined by the fact that the pronunciation of Russian vowels depends on their position with respect to the word stress, which in its turn is highly variable. On the one hand, different lemmas may have a stress on different syllables (*muká* 'flour' – *múka* 'torment'). On the other hand, in many cases the position of the stress is not constant within the paradigm of the same word (*straná* 'country, Nom, Sg' – strány 'country, Nom, Pl'). There are several hundreds of patterns that describe the distribution of stress in a paradigm. We are currently introducing these patterns in the morphological dictionary. The second challenge consists in generating natural intonation for different types of sentences. Our preliminary experiments showed that the information on the dependency structure produced by the ETAP-3 parser is instrumental in generating an adequate intonation contour.

This model is used in three environments which we will briefly comment on below: Transfer-based Machine Translation, Interlingua-based Machine Translation via UNL, and Knowledge-based Semantic Analysis.

Transfer-based MT

This is the most developed ETAP-3 option. Although there are prototypes for several language pairs, by far the most advanced is the Russian ↔ English pair. The English grammar and the dictionaries have been developed along the same lines and have approximately the same coverage as the Russian ones. Some characteristic properties of the ETAP-3 MT module are as follows.

1. In the current version of ETAP-3, the modules that process NL sentences are strictly rule-based. However, in a series of 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. Another area of research is directed towards building a hybrid parser.
2. ETAP-3 is able to offer 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 translations, 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 ...).*
3. There is an option of interactive word sense disambiguation and syntactic ambiguity resolution. The method applied consists in asking the user to identify a word sense, or a syntactic interpretation, whenever the system lacks reliable data to make the choice automatically. In lexical disambiguation, part of man-machine dialogue refers to the analysis phase, while the other part is activated during transfer.

Interlingua-based MT

This option is based on the Universal Networking Language (UNL) elaborated in the United Nations University by H. Uchida. UNL has sufficient expressive power to represent relevant information conveyed by natural languages. A consortium of several teams from different countries (Brasil, Egypt, France, India, Italy, Japan, Russia, Spain) are developing modules for translating texts from their languages to UNL and vice versa. On the basis of ETAP-3, these modules have been developed for Russian and English (Boguslavsky et al. 2000). In both cases, morphological, syntactic and deep-syntactic components of the transfer-based MT option have been re-used and supplemented with respective components relating deep-syntactic structures to UNL.

Knowledge-based Semantic Analysis

This is an ongoing project aiming at producing semantic structures of Russian texts on the basis of an ontology. Our approach is in many respects similar to the Ontological Semantics (Nirenburg, Raskin 2004) approach, although the linguistic framework is different. We are constructing an ontology of a restricted domain (news on football) in OWL2 and SWRL and relating it with ETAP-3 combinatorial dictionary. As in the case of the UNL modules, the semantic analyzer re-uses morphological, syntactic, and deep-syntactic modules of the transfer-based MT option.

References

Apresjan Ju., I. Boguslavsky, L. Iomdin, A. Lazurskij, L. Mitjushin, V. Sannikov, L. Cinman. “ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the Meaning ⇔Text Theory,” *Proceedings of the 1st International Conference on Meaning-Text Theory*. Paris, Ecole Normale Supérieure, June 16–18; 279-288, 2003.

Boguslavsky I., N. Frid, L. Iomdin, L. Kreidlin, I. Sagalova, V. Sizov. Creating a Universal Networking Language Module within an Advanced NLP System // Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), 2000, p. 83-89 (with N). (COLING 2000), 2000, p. 83-89.

Nirenburg S., Raskin V. Ontological Semantics. MIT Press, 2004.