LABORATORY 15

Laboratory of Computational Linguistics

Head of Laboratory – Dr.Sc. (Linguistics), Prof. Igor Boguslavsky
Tel.: (095) 299-49-27; E-mail: bogus@iitp.ru

The leading researchers of the laboratory include:

Full member of the Russian Academy of Sciences, Dr.Sc. (Linguistics) Jury D. Apresjan		
Dr.Sc. (Linguistics).	Vladimir Z. Sannikov	Maria S. Bulakh
Dr.	Leonid L. Iomdin	Nikolay V. Grigoriev
Dr.	Leonid G. Mitjushin	Alexander V. Lazursky
Dr.	Leonid L. Tsinman	Leonid G. Kreidlin
Dr.	Svetlana A. Grigorieva	Nadezhda E. Frid

DOMAINS OF RESEARCH

The principal focus of research is the functioning of the natural language as a means of information transmission. Fundamental research carried on in the Laboratory aims at the elaboration of a fully operative linguistic model of the "Meaning ⇔ Text" type. The model is intended to simulate the language behaviour of humans, that is, their ability to produce texts in a natural language and understand them. The computerized version of the model developed by the Laboratory is shaped as a polyfunctional multilingual processor known under the name of ETAP, which consists of morphological and combinatorial dictionaries of the working languages and various sets of rules. Ideally, the rules in combination with the dictionaries should simulate the language behavior of humans in text production and interpretation. Integrated into individual modules, they acquire a variety of applied functions. In particular, they enable the operation of a number of NLP systems designed in the Laboratory, such as English-to-Russian and Russian-to-English machine translation, generation of Russian texts from the semantic representation of an utterance in UNL, paraphrasing sentences in the given natural language, and some others. Apart from solving a number of practical problems in NLP, such systems, on the other hand, serve as an experimental testing ground, which enables the researchers to rectify linguistic descriptions and obtain entirely new linguistic knowledge from the experimental data. The demo version of the system is accessible at the following address: http://proling.iitp.ru. In 2004, the Laboratory focused on the following three directions of research: development of theoretical problems of lexicography, creation of a deeply annotated corpus of the Russian Language, and the extension and improvement of the functional potential of ETAP. In all, the research was pursued within five topical projects.

BASIC RESULTS

1. Topic: Development of a new theory and apparatus of lexical functions for computational linguistics purposes.

Lexical functions (LF) may be exemplified by semi-auxiliary verbs of the OPER1 class, which can be used to form large series of verbal-nominal paraphrases like (1) *to influence – to exert* (OPER1), (2) *influence* or *to attack – to launch* (OPER1) *an attack*. The classical theory of LFs suggests that a) verbs of this class are semanti-

cally empty: the meaning of the phrase *to exert influence* is concentrated in the noun, otherwise the synonymy of (1) could not be explained; b) the choice of verb for OPER1 is semantically unmotivated: there is no explanation why phrases like **launch influence* or **exert an attack* are ungrammatical. Our research has been able to prove that the main verbs of the OPER family are not semantically empty and that, consequently, the choice of verb for OPER1, OPER2 and a large number of its compositions (no less than 10 LFs) is semantically motivated. The motivation lies in the following general law of semantic agreement: both the LF verb and its argument always belong to one and the same class of the fundamental classification of predicates. This result creates a basis for the prognosis of the possible meaning of a given LF for a given argument; the likelihood of such prognoses amounts to 80-85 %. This prediction apparatus is very important for lexicography, learning of native and foreign languages, and natural text processing.

2. Topic: Development of the theory and dictionary of verbal government for the purposes of automatic analysis and synthesis of Russian texts.

The traditional theory of verbal government was purely syntactic. It was believed that a verb subcategorizes for a noun in a certain case (possibly with a preposition) if this verb cannot be used in basic simple texts without such a noun (cf. *rent something, depend on something* etc. In Igor Mel'čuk's Meaning ⇔ Text model, the theory of verbal government acquired a more solid semantic basis: potential syntactic dependents of a verb were considered to be strongly governed if they corresponded to semantic valences of this verb, or obligatory participants of the situation designated by it. The valences are specified by the analytic definition of the verb, cf. *A rents B from X at C for T* = 'Human A takes property B from B's owner X for the price C for the period of time T to use it in A's personal interests and on condition that B will be returned to X upon expiration of T''. Such an approach substantially enriched the view of verbal government (the verb *to rent* subcategorizes for five noun phrases and not for one). Still, verbal government was treated in a piecemeal fashion and determined for each verb separately. Our research has shown that in many cases verbs can be integrated into large classes with the same government particularities and, consequently, allow for a systematic description. The fact that a verb belongs to a particular syntactic class is determined by its belonging to a semantic class/subclass of the fundamental classification of predicates.

3. Topic: Morpho-syntactic and lexico-semantic annotation of Russian texts.

The first deeply annotated corpus of Russian texts has been created. Every sentence of the 25-thousand sentence strong corpus is supplied with a full morphological and syntactic annotation. Morphological tagging includes information of the lemma (dictionary form) of every word, its part of speech, and the complete list of inflectional grammatical features peculiar to this word. Syntactic tagging is presented as a dependency tree structure, which is built using the techniques and means of the ETAP-3 linguistic processor. A metalanguage for semantic annotation of the corpus has been developed. The metalanguage includes generic and specific descriptors (semantic features) for object names (about 90), generic and specific descriptors for predicates (over 100), and semantic roles (about 50). For the first release of the semantically tagged corpus, relevant sets of descriptors have been assigned to 3,000 words. A highly efficient and convenient software package for semi-automatic text annotation has been developed. First, every sentence is passed through the analyzing module of; later, the produced syntactic structure is edited by a linguist who works in a specially designed software environment. In the nearest future, access to the deeply annotated corpus will be available through a public Internet site.

4. Topic: Development of an interactive system of lexical ambiguity resolution for machine translation and other applications.

The main problem of natural text analysis is a high level of lexical unit ambiguity. A new module of the ETAP-3 linguistic processor has been created, which enables a resolution of such ambiguity through dynamic human/computer interaction. The module of interactive lexical disambiguation, activated at such points of the analysis algorithm where the linguistic processor lacks data to determine which of the potential readings of a word used in a text is relevant, offers the system user several options to choose from. The user is guided by his linguistic competence, extralinguistic data and intuition when choosing the option. Upon receiving the user's response, the processor resumes the processing of the current sentence. As shown by a broad series of experiments with the newly developed module, the implemented technique ensures a high degree of ambiguity resolution which cannot be achieved by purely automatic means.

5. Topic: Development of a module of syntactic and semantic analysis for the system of multilingual communication in the Internet based on the Universal Networking Language (UNL).

The system of multilingual communication in the Internet is jointly developed by a large international consortium that consists of research groups from 14 countries. The system is based on an artificial language of meaning representation (UNL) which acts as an interlingua for all working languages of the system. The research has resulted in the creation of s set of rules intended to convert Russian texts into UNL hypergraphs. These rules constitute an important part of the UNL-module which is being developed as a separate subsystem of the ETAP-3 linguistic processor. Together with the Russian-UNL dictionary, the rules enable an efficient construction of UNL texts and, eventually, the functioning of an interface between Russian and the remaining languages of the system. Experiments on harmonization of semantic and syntactic analysis rules for different languages were carried out in cooperation with foreign partners. The development of an integrated corpus of UNL texts has been started. This corpus will serve as basis for the creation of UNL modules for new languages.

AWARDS

- Yuri D. Apresjan was awarded with the Vladimir Dal Gold Medal for the "New explanatory dictionary of Russian synonyms" in three volumes.

GRANTS FROM:

- **Russian Foundation of Basic Research (No. 02-06-80085):** «Development of an interactive system of lexical ambiguity resolution for machine translation and other applications».
- **Russian Foundation of Basic Research (No. 02-06-80106):** «Development of the theory and dictionary of verbal government for the purposes of automatic analysis and synthesis of Russian texts».
- **Russian Foundation of Humanitarian Research (No. 03-04-00046a):** «Development of a new theory and apparatus of lexical functions for computational linguistics purposes».
- **Russian Foundation of Humanitarian Research (No. 04-04-00263):** «Unsolved problems of Russian Syntax».

- **Russian Foundation of Basic Research (No. 04-06-80148):** «Development of a module of syntactic and semantic analysis for the system of multilingual communication in the Internet based on the Universal Networking Language (UNL)».
- **Russian Foundation of Basic Research (No. 04-07-90179):** «Development of a corpus of Russian text supplied with morpho-syntactic annotation».
- **State Contract within the program "Philology and information science" (No. 10002-251/OIFN-02/241-096/280504-384):** «Morph-syntactic and lexico-semantic annotation of a subarray of a balanced text corpus».
- **Contract from the Ministry of Industry and Science of Russia (No. 31):** «Models and algorithms of information interaction in the domains of genetics, linguistics, and color vision».

WORK WITH YOUNG SCIENTISTS

In 2004, laboratory staff members taught at the Moscow State University and the Moscow Institute of Foreign Languages (Yuri D. Apresjan), Russian State Humanitarian University (Leonid L. Iomdin) and Madrid Polytechnic University (Igor M. Boguslavsky). Yuri D. Apresjan and Leonid L. Iomdin supervized the research of postgraduate students within the research domains of the laboratory

PUBLICATIONS IN 2004

Published papers

1. Апресян В.Ю., Апресян Ю.Д., Богуславская О.Ю., Крылова Т.В., Левонтина И.Б., Урысон Е.В. и др. М., Школа «Языки славянской культуры», 2004, LXIX, 1417 с. Новый объяснительный словарь синонимов русского языка. 2-е издание, исправленное и дополненное. Под общим руководством академика Ю. Д. Апресяна.
2. Апресян Ю.Д. Акциональность и стативность как сокровенные смыслы (охота на *оказывать*) // Сокровенные смыслы. Сборник статей в честь Н. Д. Арутюновой. Главный редактор Ю. Д. Апресян. М., Школа «Языки славянской культуры», 2004, 13-33.
3. Апресян Ю.Д. Интерпретационные глаголы: семантическая структура и свойства // Русский язык в научном освещении, «Языки славянской культуры», 2004, № 7, 5-22.
4. Апресян Ю.Д. О семантической непустоте и мотивированности глагольных лексических функций // ВЯ, «Наука», 2004, № 4, 3-18.
5. Апресян Ю.Д. О семантической непустоте и мотивированности лексических функций глаголов // Проблемы русской лексикографии. Тезисы докладов международной конференции "Шестые Шмелевские чтения". 24-26 февраля 2004 г. М., ИРЯ им. В. В. Виноградова РАН, 2004, 4-10.
6. Апресян Ю.Д. Принципы организации центра и периферии в лексике и грамматике // Типологические обоснования в грамматике. К 70-летию профессора В. С. Храковского. М.: Языки славянской культуры, 2004, 20-35.
7. Богуславский И.М. Как значение предложения складывается из значений слов. // E. F. Quero Gervilla, A. Salmeron Vilchez (eds.) III Jornadas andaluzas de eslavistica. Granada, 2004, 32-33.
8. Богуславский И.М. Целое – часть – признак. // Сокровенные смыслы. Сборник статей в честь Н. Д. Арутюновой. Главный редактор Ю. Д. Апресян. М.: Языки славянской культуры, 2004, 44-53.
9. Иомдин Л.Л. Идея и цель: об одном типе русских связочных предложений // Сокровенные смыслы. Сборник статей в честь Н. Д. Арутюновой. Главный редактор Ю. Д. Апресян. М.: Языки славянской культуры, 2004, 418-425.

10. Иомдин Л.Л. Лексикографический портрет наречия что-то // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог'2003. М.: Наука, 2004, 246-251.
11. Иомдин Л.Л. Уроки машинного перевода для детей и взрослых. // Лингвистика для всех. Зимняя лингвистическая школа - 2004. М., НИИРО, 2004. 56-68.
12. Санников В.З. О значении союза *пускай/пусть*. // Отцы и дети Московской лингвистической школы. Памяти Владимира Николаевича Сидорова. М., Институт русского языка, 2004. 239-245.
13. Bulakh M. Color terms of Modern South Arabian Languages: a Diachronic Approach. // Orientalia. Papers of the Oriental Institute, Vol. V. Babel und Bibel 1. Moscow, 2004, 269-292.
14. Iomdin L., Boguslavsky I., Sizov V. Multilinguality in ETAP-3. Reuse of Linguistic Resources // Proceedings of the Workshop "Multilingual Linguistic Resources". 20th International Conference on Computational Linguistics. 7-14.

Papers in print

1. Апресян Ю.Д. Два принципа и два понятия системной лексикографии // Сборник в честь Т.М. Николаевой. 2 п. л.
2. Апресян Ю.Д. О Московской семантической школе // ВЯ, «Наука», 2005, № 1. 2 п. л.
3. Апресян Ю.Д. Правила взаимодействия значений и словарь // Русский язык в научном освещении, «Языки славянской культуры», 2005, № 8. 2 п. л.
4. Апресян Ю.Д. Трактовка вида в словаре: правила, тенденции, лексикализация // Сборник в честь проф. Лемана. Hamburg. 1 п. л.
5. Апресян Ю.Д., Апресян В.Ю., Бабаева Е.Э., Богуславская О.Ю., Иомдин Б.Л., Крылова Т.В., Левонтина И.Б., Санников А.В., Урысон Е.В. Основания системной лексикографии // Языковая картина мира и системная лексикография. Коллективная монография под общим руководством акад. Ю. Д. Апресяна. Авторы: Около 40 п. л. М.: Языки славянской культуры.
6. Апресян Ю.Д., Иомдин Л.Л., Санников А.В., Сизов В.Г. Семантическая разметка в глубоко аннотированном корпусе русского языка // Материалы конференции по корпусной лингвистике. СПб: Санкт-Петербургский ун-т.
7. Богуславский И.М. Валентности кванторных слов. // Сборник "Квантификативный аспект языка". М.
8. Иомдин Л.Л. Новые наблюдения над синтаксисом русских фразем // Obecność. Uniwersytet w Białymstoku.
9. Санников В.З. Иллокутивное употребление или синтаксический эллипсис? // Русский язык в научном освещении. М., Институт русского языка РАН, № 9, 2005.
10. Boguslavsky I. Some controversial issues of UNL: linguistic aspects // Proceedings of the 2nd International workshop on UNL and other interlinguas. Mexico, 2005.
11. Boguslavsky I., Cardenosa J., Gallardo C., Iraola L. The UNL initiative: an Overview // Proceedings of CICLING-2005, Mexico, 2005.
12. Boguslavsky I., Iomdin L. Moscow semantic school. Berlin, 2005.
13. Boguslavsky I., Iomdin L., Mityushin L.et al. Interactive Resolution of Intrinsic and Translational Ambiguity in a Machine Translation System // Proceedings of CICLING-2005, Mexico, 2005.