Institute for Information Transmission Problems

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	Nikolay V. Grigoriev
Dr.	Leonid L. Iomdin	Alexander V. Lazursky
Dr.	Leonid G. Mitjushin	Irina E. Kayali
Dr.	Leonid L. Tsinman	Leonid G. Kreidlin
Dr.	Svetlana A. Grigorieva	Nadezhda E. Frid
	Maria S. Bulakh	

DOMAINS OF RESEARCH

The principal focus of laboratory research is the functioning of the natural language as a means as a means of information transmission Fundamental research carried on in the Laboratory aims at the elaboration of a full 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 poly-functional multilingual processor known under the name of ETAP. ETAP 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 behaviour of humans in text production and interpretation. Integrated into specific modules, they acquire a variety of applied functions. In particular, they enable the operation 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. In 2003, the Laboratory was concerned with the extension and improvement of the functional potential of ETAP. The demo version of the system is accessible at the following address: http://proling.iitp.ru.

BASIC RESULTS

1) Much effort was devoted to the replenishment and development of the combinatorial dictionaries of Russian and English.

– The number of entries has grown to 75,000 items in the Russian dictionary, and 68,000 entries in the English dictionary, which somewhat exceeds the average size of large traditional all-purpose bilingual dictionaries.

– Lexicographic information in various zones of the dictionaries was expanded and updated. Fundamental changes were introduced into a) the entries of polysemic words, b) government patterns, c) lexical functions zones. The expansion and updates were motivated, in particular, by the experimental data drawn from the functioning of applied systems constructed by the Laboratory.

2) Work on the replenishment and improvement of morphological dictionaries of Russian and English was carried on.

– The dictionaries were replenished with the material of new texts.

– Introduction of composites (of the type *quasi*) and the components of compound words was continued.

3) Studies on the topic "Elaboration of the theory and dictionary of verbal government for the purpose of automatic parsing and generation of Russian texts" were continued. The following results were obtained:

– New analytical definitions of verbs were elaborated.

– A detailed semantic classification of verbs was proposed within the fundamental classification of predicates. Every class of Russian verbs was matched with a similar class of English verbs.

– The previous version of the fundamental classification of predicates was substantially rectified and expanded.

– A new classification of semantic roles of valencies of predicate words was proposed.

4) Research into the theory of linguistic functions was continued. The following main results were obtained:

– A new version of the theory of lexical functions was proposed.

– Formal (systematic) definitions were constructed for all lexical functions.

– A number of new lexical functions were described.

– Representative lists of lexical functions were created (depending on the function, each lists contains from several words to several hundred argument words for which all values of every function are specified).

5) The construction of the Russian subsystem for the programme of multilanguage communication on the basis of the Universal Networking Language (UNL) was continued. The programme, jointly developed by a large international consortium under the aegis of the UN, is aimed at providing the Internet users with the facilities for receiving and disseminating information in their mother tongues. The system is accessible at the site www.undl.org. The following results were obtained.

– The second version of the Russian-to-UNL translation module was developed.

– The module for UNL-to-Russian translation was drastically improved. This module is accessible at the address www.unl.ru.

– The first version of UNL modules for English was developed.

– All UNL modules of the ETAP system were integrated with the EditorUNL, system, developed by the laboratory's partners from Madrid Technical University.

6) Work on creating a syntactically annotated corpus of Russian texts was carried on.

– Over 6,000 sentences were annotated for the corpus.

– The software of the annotator's workbench was improved and updated.

– A fragment of a syntactically annotated corpus was developed within the framework of a fundamental research program of the Russian Academy of Sciences target towards the creation of the Russian National Corpus.

7) Improvements were made for the combined parsing algorithm in which syntactic rules dynamically interact with the specially designed statistical module, which assigns weights to the hypothetical syntactic relations on the basis of the syntactically

annotated corpus. A candidate dissertation was prepared within this research framework.

8) The research and development work on the interactive system of lexical ambiguity resolution for machine translation and other NLP applications was continued. The system is based on a dynamic interaction of the algorithm with the intuitions of the human user who has no linguistic expertise. On receiving the user's answers to some simple questions posed to him the system enables a hundred percent resolution of lexical ambiguity. With the help of this system some 20,000 entries in the combinatorial dictionary were processed for which the structure of their polysemy was established and diagnostic descriptions for various meanings were introduced.

9) The interactive lexical ambiguity resolution system was integrated with the UNL module.

GRANTS FROM:

- **Russian Foundation of Basic Research (No. 01-07-90405):** "Creation of an Annotated Corpus of Russian Texts (second release)".
- **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 the Dictionary of Verbal Valencies 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".

PUBLICATIONS IN 2003

1. Апресян Ю.Д. Системность лексики: семантические парадигмы и семантические альтернации // Etudes linguistiques romano-slaves offertes à Stanislaw Karolak, Oficyna Wydawnicza "Edukacja", 2003. P. 35-47.
2. Апресян Ю.Д. Фундаментальная классификация предикатов и системная лексикография. Грамматические категории: иерархии, связи, взаимодействие // Материалы международной конференции. СПб.: Наука, 2003. С. 7-21.
3. Апресян Ю.Д. Акциональность и стативность как сокровенные смыслы (охота на оказывать) // Сокровенные смыслы. Сборник статей в честь Н.Д. Арутюновой. М.: Языки славянской культуры, 2004.
4. Апресян Ю.Д. Принципы организации центра и периферии в лексике и грамматике (in print).
5. Апресян Ю.Д. Интерпретационные глаголы: семантическая характеристика и свойства // Русский язык в научном освещении. Языки славянской культуры (in print).
6. Апресян Ю.Д., Цинман Л.Л. Формальная модель перифразирования предложений для систем переработки текстов на естественных языках // Русский язык в научном освещении. 2002. № 2 (4). С. 102-146. М.: Языки славянской культуры, 2002 (реально вышел в 2003 г.).
7. Апресян Ю. Д., Апресян В.Ю., Богуславская О.Ю. и др. Новый объяснительный словарь синонимов русского языка. Третий выпуск. М.: Языки славянской культуры, 2003. С. I-LXV, 1-557.

8. Богуславский И. Замечания об актантной структуре адвербиальных дериватов // Die het kleine eert, is het grote weerd. Pegasus Oost-Europese Studies 1, Uitgeverij Pegasus, Amsterdam, 2003, 23-40.
9. Богуславский И.М. Часть – Целое – Признак: заметки о сфере действия кванторных слов (in print).
10. Богуславский И.М., Иомдин Л.Л., Сизов В.Г., Чардин И.С. Использование размеченного корпуса текстов при автоматическом синтаксическом анализе // Международная конференция «Когнитивное моделирование в лингвистике». Сборник докладов. Варна, 2003. С. 39-48.
11. Иомдин Л.Л. Идея и цель: об одном типе русских связочных предложений // Сокровенные смыслы. Сборник статей в честь Н.Д. Арутюновой. М.: Языки славянской культуры, 2004.
12. Иомдин Л.Л. Большие проблемы малого синтаксиса // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог'2003. Протвино, 2003. С. 216-222.
13. Иомдин Л.Л. Новые наблюдения над синтаксисом русских фразем // Obecność. Uniwersytet w Białymstoku. 2003 (in print).
14. Санников В.З. Русская языковая шутка. М.: Изд-во «Аграф», 2003.
15. Apresjan Ju.D. Principles of Systematic Lexicography. Lexicography and Natural Language Processing // A Festschrift in Honour of B.T.S. Atkins. Marie-Hélène Corréard, 91-104, 2002 (реально вышел в 2003 г.)
16. Apresjan Ju.D., Boguslavsky I., Iomdin L., Tsinman L. Lexical Functions in ETAP-3 // Proceedings of the First International Conference on the "Meaning – Text" Theory. Сайт в Интернете http://mtt2003.linguist.jussieu.fr/actes/index.en.html.
17. Apresian J., Boguslavsky I., Lazursky A., Sannikov V., Sizov V., Tsinman L. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT // MTT 2003, First International Conference on Meaning – Text Theory. Ecole Normale Superieure, Paris, June 16-18, 2003. P. 279-288.
18. Boguslavsky I. On the Passive and Discontinuous Valency Slots // Proceedings of the 1st International Conference on Meaning-Text Theory. Ecole Normale Supérieure, Paris, June 16-18, 2003.
19. Boguslavsky I. Breaking the language barrier in the Internet // Die Kosten der Mehrsprachigkeit. Globalisierung und sprachliche Vielfalt. Rudolf de Cillia, Hans-Jürgen Krumm, Ruth Wodak (Hrsg.). Verlag der Österreichischen Akademie der Wissenschaften, Wien, 2003, 143-145.
20. Boguslavsky I., Iomdin L., Sizov V. Interactive enconversion by means of the ETAP-3 system // Proceedings of the International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies. Alexandria, 2003.
21. Bulakh M. Etymological notes on the Akkadian Colour Terms. // Orientalia, Papers of the Oriental Institute, Vol III: Studia Semitica, 3-17, ed. L. Kogan. Oriental Institute: Moscow, 2003.
22. Iomdin L. Natural Language Processing as a Source of Linguistic Knowledge. // Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. Las Vegas, June 23-26 2003. P. 68-74.
23. Iomdin L. Purpose and Idea: a Lesson Drawn from Machine Translation. // MTT 2003. First International Conference on Meaning – Text Theory. Ecole Normale Superieure, Paris, June 16-18 2003. P. 269-278.