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

DOMAINS OF RESEARCH

The Laboratory is concerned with the study of language as a means of information transmission. Fundamental research carried on in the Laboratory aims at the elaboration of a full operative "Meaning ⇔ Text" type of linguistic model. The model is intended to simulate the language behaviour of humans, that is, their ability to produce and understand texts in a natural language. 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 certain modules, they acquire an applied function. Namely, they make possible 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 allows the researchers to rectify certain details of linguistic descriptions and sometimes even derive totally new linguistic knowledge from the experimental data. In 2002 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 development and replenishment of the combinatorial dictionaries of Russian and English.

– The number of entries has grown to 65,000 items in each dictionary, which corresponds to 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 up-

dates were motivated, in particular, by the experimental data drawn from the functioning of applied systems constructed by the Laboratory.

– A module for semi-automatic introduction of phraseological units into the combinatorial dictionaries was developed.

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

– The dictionaries were replenished with geographical and proper names.

– 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 parsing and automatic synthesis of Russian texts” were commenced. The following results were obtained:

– A new lexico-semantic theory of verbal government was elaborated. It can be summed up in the following points: a) verbal government is partly motivated semantically and partly unmotivated; b) those governing properties of verbs which can be accounted for by semantic considerations are determined by two factors – the systemic senses of the type ‘action’, ‘process’, ‘state’, ‘property’, ‘actionality’, ‘purpose’ etc., making part of the lexicographic definitions of the respective predicate lexemes, and by those definitions themselves; c) unmotivated governing properties of lexemes are rooted in the "personal" lexical properties of the verbs.

– A list of verbs to be included in the dictionary was compiled counting some 6000 items.

4) A special module for paraphrasing Russian sentences on the basis of the lexical functions apparatus was developed; the following results were obtained in the course of this work:

– A new apparatus of lexical functions was constructed consisting of the new system of definitions, the zone of lexical functions in the combinatorial dictionaries and three sets of rules: rules of identification of lexical functions in the processed sentence, rules of reducing sentences to their canonical forms and rules of paraphrasing proper. This apparatus was integrated into ETAP as a separate module.

– The system of paraphrasing was debugged, and a large series of computer experiments were conducted.

– The examination of the experimental data was commenced which allowed not only to introduce a number of local improvements in the linguistic knowledge base, but also to make considerable headway in the understanding of the fundamental properties of language.

5) Much work was done on constructing the Russian subsystem for the programme of multi-language communication on the basis of the Universal Networking Language (UNL). This programme, which is being 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 first version of Russian-to-UNL translation module was developed.

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

– The existing specifications of UNL were analyzed and a number of improvements were proposed.

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

– 4 000 sentences in the corpus were annotated.

– The software of the complex "The working bench of the annotator" was perfected: a program was written which either corrects itself certain mistakes found in the text or detects formal inconsistencies in the morpho-syntactic structures and supplies the operator with a list of mistakes.

– A fragment of the corpus was handed over to the YANDEX company for joint effort at developing an information retrieval system and creating an open telecommunication access to the corpus via Internet.

7) An original combined parser was constructed and integrated into the ETAP linguistic processor. It allows parsing rules to dynamically interact in the process of ambiguity resolution with the specially designed statistical module which assigns certain weights to the hypothetical syntactic relations on the basis of the syntactically annotated corpus.

8) The first version of an innovative interactive system of lexical ambiguity resolution for machine translation and other NLP applications was constructed. 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 (choose a synonym of the given word, or its paraphrase, or its simplified definition and the like) the system ensures 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) Linguistic means of lexico-grammatic ambiguity resolution were worked out.

10) A program for identifying and tagging nominal groups in the text which designate persons was written.

11) Support for accessing symbols of various code tables by ETAP was ensured.

12) A new version of the HELP system for ETAP was elaborated.

GRANTS FROM:

- **Russian Foundation of Basic Research (No. 01-06-80453):** "Development of a Compound Parsing Algorithm for the Linguistic Processor ETAP-3".
- **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".

PUBLICATIONS IN 2002

1. Апресян Ю.Д. Об одной закономерности устройства семантических систем // Проблемы семантического анализа лексики. Тезисы докладов международной конференции. М., 2002. С. 6-9.
2. Апресян Ю.Д. Взаимодействие лексики и грамматики: лексикографический аспект // Русский язык в научном освещении. 2002. № 3. С. 10-29.

3. Апресян Ю.Д. Новый объяснительный словарь синонимов русского языка: ход работы и результаты // Вестник Российского гуманитарного научного фонда. 2002. № 3. С. 87-99.
4. Апресян Ю.Д., Цинман Л.Л. Формальная модель перифразирования предложений для систем переработки текстов на естественных языках // Русский язык в научном освещении. 2003. № 4.
5. Богуславский И.М. «Сандхи» в синтаксисе: загадка *уже не* // ВЯ. 2002. № 5. С. 19-37.
6. Богуславский И.М., Григорьев Н.В., Григорьева С.А., Иомдин Л.Л. Разработка синтаксически размеченного корпуса русского языка // Доклады научной конференции "Корпусная лингвистика и лингвистические базы данных". С.-Пб.: изд-во С.Петербургского университета, 2002. С. 40-50.
7. Иомдин Л.Л. Уроки русско-английского (из опыта работы системы машинного перевода) // Труды Международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. М.: Наука, 2002. Т. 2. С. 234-244.
8. Иомдин Л.Л., Сизов В.Г., Цинман Л.Л. Использование эмпирических весов при синтаксическом анализе // Обработка текста и когнитивные технологии. Казань: Отечество, 2002. № 6. С. 64-72.
9. Фрид Н.Е. Употребление настоящего исторического и прошедшего времен в спонтанной устной речи // Международная школа по лингвистической типологии и антропологии. Материалы лекций и семинаров. М.: РГГУ, 2002. С. 268-269.
10. Apresjan Ju.D. Principles of Systematic Lexicography // Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins. Marie-Hélène Corréard (ed), Euralex 2002. P. 91-104.
11. Apresjan Ju.D., Boguslavsky I.M., Iomdin L.L., Tsinman L. L. Lexical Functions in NLP: Possible Uses. – In: Computational Linguistics for the New Millenium: Divergence or Synergy? Proceedings of the International Symposium held at the Ruprecht-Karls-Universität Heidelberg, 21-22 July 2000. Manfred Klenner / Henriëtte Visser (eds.) Frankfurt am Main, 2002. P. 55-72.
12. Boguslavsky I., Chardin I., Grigorjeva S., Iomdin L. et al. Development of a dependency treebank for Russian and its possible applications in NLP // Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), v. III, Las Palmas. P. 852-856.
13. Boguslavsky I. Some lexical issues of UNL // Proceedings of the First International Workshop on UNL, other interlinguas and their applications, Las Palmas, 2002. P. 19-22.
14. Iomdin L., Sizov V., Tsinman L. Utilisation des poids empiriques dans l'analyse syntaxique: une application en Traduction Automatique // META. 2002. V. 47. No. 3. P. 351-358.
15. Апресян Ю.Д. Остановка движения как симптом внутреннего состояния: синонимический ряд *замереть* // Сборник в честь В. Н. Сидорова (в печати).
16. Апресян Ю.Д. Системность лексики: семантические парадигмы и семантические альтернации // В сборник в честь С. Кароляка (в печати).
17. Апресян Ю.Д. Акциональность и стативность как сокровенные смыслы (охота на *оказывать*) // Сборник статей к 70-летию Н.Д. Арутюновой. М.: Языки русской культуры, 2003 (в печати).
18. Апресян Ю.Д. Трактовка вида в словаре: правила, тенденции, лексикализация // Сборник в честь проф. Лемана, Гамбург (в печати).

19. Апресян Ю.Д. Об одной закономерности устройства семантических систем // Пятые Шмелевские чтения (в печати).
20. Апресян Ю.Д. Принципы организации центра и периферии в лексике и грамматике // Сборник в честь В.С. Храковского (в печати).
21. Богуславский И.М. Замечания об актантной структуре адвербиальных дериватов (в печати).
22. Богуславский И.М. Часть – Целое – Признак: заметки о сфере действия кванторных слов (в печати).
23. Иомдин Л.Л. Идея и цель: об одном типе русских связочных предложений // Сборник статей к 70-летию Н.Д. Арутюновой. М.: Языки русской культуры, 2003 (в печати).
24. Apresjan Ju.D., Boguslavsky I.M., Iomdin L.L., Tsinman L.L. Lexical Functions in ETAP-3 // Сборник в честь 70-летия И.А. Мельчука (в печати).