ЛАБОРАТОРИЯ № 15

Лаборатория компьютерной лингвистики

Заведующий лабораторией: д.филол.н., Богуславский Игорь Михайлович
Тел.: (095) 299-49-27; Email: bogus@iitp.ru

Ведущие ученые лаборатории:

академик, д.филол.н.	Апресян Ю. Д.	Булах М. С.
д.филол.н.	Санников В. З.	Григорьев Н. В.
к.филол.н.	Григорьева С. А.	Крейдлин Л. Г.
к.филол.н.	Иомдин Л. Л.	Лазурский А. В.
к.ф.-м.н.	Митюшин Л. Г.	Фрид Н. Е.
к.ф.-м.н.	Цинман Л. Л.	

НАПРАВЛЕНИЯ ИССЛЕДОВАНИЙ

Основной научной проблематикой лаборатории является функционирование естественного языка в качестве средства передачи информации. Фундаментальные исследования, проводимые в лаборатории, направлены на разработку полной действующей формальной модели языка типа "Смысл ⇔ Текст". Модель должна имитировать языковое поведение человека, т.е. его способность производить тексты на естественном языке и понимать их. Компьютерная версия модели, разрабатываемая в Лаборатории, имеет вид полифункционального многоязычного процессора, известного под условным наименованием ЭТАП. ЭТАП состоит из морфологических и комбинаторных словарей рабочих языков и наборов правил. В идеале правила во взаимодействии со словарями должны имитировать языковое поведение человека при производстве и понимании тестов. Объединенные в определенные модули, они приобретают и прикладную функцию, а именно, обеспечивают функционирование ряда построенных в Лаборатории систем переработки текстов, таких, как машинный перевод, порождение русских текстов по смысловому заданию на языке UNL, перифразирование предложений на данном естественном языке и т. п. Будучи инструментом решения ряда практических задач в области переработки текстов на естественных языках, такие системы, с другой стороны, служат экспериментальным полигоном для корректировки лингвистических описаний и получения принципиально новых лингвистических знаний. В 2003 году работа Лаборатории была направлена на расширение и совершенствование функциональных возможностей системы ЭТАП. Демонстрационная версия системы доступна по адресу http://proling.iitp.ru.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1) Продолжалась работа над пополнением и развитием комбинаторных словарей русского и английского языков.

– Словники этих словарей доведены до объема около 75000 единиц (русский словарь) и 68000 единиц (английский словарь), что соответствует объему крупных традиционных двуязычных словарей общего назначения.

– Лексикографическая информация в разных зонах словарей была пополнена и откорректирована. Наиболее существенной корректировке подверглись а) словарные статьи многозначных слов, б) модели управления, в) зона лексических функций. Основанием для пополнения и корректировки послужили, в ча-

стности, экспериментальные материалы, полученные в ходе функционирования разработанных в Лаборатории прикладных систем.

2) Продолжалась работа над пополнением и развитием морфологических словарей русского и английского языков.

– Словари пополнялись за счет новых текстов.

– Продолжалась работа по введению в морфологические словари композитов (типа квази- и quasi-) и компонентов сложных слов.

3) Продолжены исследования по теме "Разработка теории и словаря глагольного управления для целей автоматического анализа и синтеза текстов на русском языке". Получены следующие результаты:

– Разработаны новые аналитические толкования глаголов.

– Разработана подробная семантическая классификация глаголов внутри фундаментальной классификации предикатов.

– Каждый класс русских глаголов поставлен в соответствие аналогичному классу глаголов английского языка.

– Была серьезно уточнена и дополнена предыдущая версия фундаментальной классификации предикатов и построена новая классификация семантических ролей актантов предикатных слов.

4) Продолжены исследования по теории лексических функций. Получены следующие результаты:

– Разработана новая версия теории лексических функций.

– Построены формальные (системные) определения всех лексических функций.

– Описан ряд новых лексических функций.

– Составлены представительные списки лексических функций (от нескольких слов до нескольких сотен слов-аргументов с указанием всех значений каждой функции для данного аргумента).

5) Продолжалась работа по созданию русской подсистемы для программы многоязычной коммуникации на базе Универсального сетевого языка (UNL). Эта программа, разрабатываемая большим международным консорциумом под эгидой ООН, ставит своей целью обеспечить для пользователей Интернета возможность получать и распространять информацию на родном языке. Познакомиться с этой программой можно на сайте www.undl.org. Получены следующие результаты:

– Разработана вторая версия модуля перевода с русского языка на UNL.

– Проведено совершенствование модуля перевода с языка UNL на русский. Этот модуль доступен по адресу www.unl.ru.

– Разработана первая версия UNL-модулей для английского языка.

– Проведена интеграция всех UNL-модулей системы ЭТАП с системой EditorUNL, разработанной партнерами из Мадридского технического университета.

6) Продолжалась работа по построению синтаксически размеченного корпуса русских текстов.

– Проведена разметка более чем 6000 предложений корпуса.

– Усовершенствовано программное обеспечение комплекса "Рабочее место аннотатора".

– Разработан фрагмент синтаксически размеченного корпуса в рамках программы фундаментальных исследований РАН по созданию Русского национального корпуса.

7) Усовершенствован комбинированный алгоритм синтаксического анализа, в котором синтаксические правила динамически взаимодействуют со специально разработанным статистическим модулем, который приписывает веса гипотети-

ческим синтаксическим связям на основе данных синтаксически размеченного корпуса текстов. Подготовлена к защите кандидатская диссертация на эту тему.

8) Продолжились работы по построению интерактивной системы разрешения лексической неоднозначности для машинного перевода и других приложений. Система основана на динамическом взаимодействии алгоритма анализа с интуицией пользователя-человека, не имеющего никакой специальной лингвистической подготовки. Получив от пользователя ответы на предложенные ему простые вопросы, система обеспечивает стопроцентное различение лексической неоднозначности. С помощью этой системы обработано 20000 слов комбинаторного словаря, для которых выявлена структура полисемии и построены диагностические описания.

9) Проведена интеграция интерактивной системы разрешения лексической неоднозначности с модулем UNL.

ГРАНТЫ:

- **Российский фонд фундаментальных исследований (№ 01-07-90405):** "Создание аннотированного корпуса русских текстов (вторая очередь)".
- **Российский фонд фундаментальных исследований (№ 02-06-80085):** "Разработка интерактивной системы разрешения лексической неоднозначности для машинного перевода и других приложений".
- **Российский фонд фундаментальных исследований (№ 02-06-80106):** "Разработка теории и словаря глагольного управления для целей автоматического анализа и синтеза текстов на русском языке".
- **Российский гуманитарный научный фонд (№ 03-04-00046а):** "Разработка новой теории и аппарата лексических функций для целей компьютерной лингвистики".

ПУБЛИКАЦИИ В 2003 г.

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. (В печати).
5. Апресян Ю.Д. Интерпретационные глаголы: семантическая характеристика и свойства // Русский язык в научном освещении. Языки славянской культуры (в печати).
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. (В печати).
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 (в печати).
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.