ЛАБОРАТОРИЯ № 15

Лаборатория компьютерной лингвистики

Заведующий лабораторией – д.филол.н., профессор
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 И ОСНОВНЫЕ РЕЗУЛЬТАТЫ

Основной научной проблематикой лаборатории является функционирование естественного языка в качестве средства передачи информации. Фундаментальные исследования 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 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. Демонстрационная версия системы ЭТАП доступна по адресу http://proling.iitp.ru. В 2004 году работа Лаборатории разворачивалась по трем основным направлениям: развитие теоретических проблем лексикографии, построение глубоко аннотированного корпуса русского языка и совершенствование функциональных возможностей системы ЭТАП. Всего в лаборатории велись исследования по пяти научным темам.

1. Разработка новой теории и аппарата ЛФ для целей компьютерной лингвистики. Примером лексической функции (ЛФ) являются полуслужебные глаголы типа OPER1, с помощью которых образуются большие серии глагольно-именных перифраз; ср. *влиять – оказывать* (OPER1) *влияние, атаковать – производить* (OPER1) *атаку* и т. п. Классическая теория ЛФ утверждает, что а) глаголы этого класса являются семантически пустыми: значение сочетания *оказывать влияние* сосредоточено в существительном, в противном случае не

было бы синонимичности *влиять = оказывать влияние*; б) выбор такого глагола на роль OPER1 семантически не мотивирован: нельзя объяснить, почему не говорят *производить влияние*, *оказывать атаку.* В ходе исследования было доказано, что основные глаголы семейства не являются семантически пустыми и что поэтому выбор глагола на роль ЛФ типа OPER1, OPER2 и большого числа их композиций (не менее 10 ЛФ) семантически мотивирован. Он объясняется следующим общим законом семантического согласования: и ЛФ-глагол, и его аргумент всегда принадлежат к одному и тому же классу фундаментальной классификации предикатов. Тем самым создается основа для прогнозов о возможном значении данной ЛФ от данного аргумента, достоверность которых составляет 80-85%. Этот прогностический аппарат имеет большое значение для лексикографии, обучения родному и иностранному языкам и для систем переработки текстов на естественных языках.

2. Разработка теории и словаря глагольного управления для целей автоматического анализа и синтеза текстов на русском языке. Традиционная теория глагольного управления была чисто синтаксической: считалось, что глагол V_i управляет существительным в определенном падеже, возможно, с предлогом, если в простейших текстах он не может употребляться без него; ср. *арендовать что-л., зависеть от кого-л*. В модели «Смысл ⇔ Текст» И. А. Мельчука теория глагольного управления приобрела более твердое семантическое основание: управляемыми стали считаться такие потенциальные синтаксические зависимые данного глагола, которым соответствуют его семантические актанты, или обязательные участники обозначаемой им ситуации. Они задаются аналитическим толкованием глагола; ср. *А арендует В у Х-а за С на Т* = ‘Человек А берет за плату С собственность В у ее владельца Х на срок Т для использования в собственных интересах, с условием, что она будет возвращена Х-у после истечения Т’. Это существенно обогатило представления о глагольном управлении (*арендовать* управляет не одной, а пятью именными группами). Однако управление в целом продолжало трактоваться «штучно» - для каждого глагола по отдельности. В ходе исследования было показано, что во многих случаях глаголы объединяются в большие классы с совпадающими управляющими свойствами и, следовательно, допускают системное описание, причем принадлежность к синтаксическому классу определяется тем, к какому семантическому классу и подклассу фундаментальной классификации предикатов принадлежит данный глагол.

3. Морфо-синтаксическая и лексико-семантическая разметка русских текстов. Создан первый глубоко аннотированный корпус русских текстов общим объемом в 25 тысяч предложений. Каждое предложение снабжено полной морфологической и синтаксической разметкой. Морфологическая разметка состоит из информации о лемме (словарной форме) каждого слова, его частеречной принадлежности и полного списка словоизменительных грамматических характеристик, присущих данному слову. Синтаксическая разметка представляет собой синтаксическую структуру в виде дерева зависимостей, построенную с использованием установок и средств лингвистического процессора ЭТАП-3. Разработан метаязык для семантической разметки корпуса. В него входят родовые и видовые дескрипторы для предметных имен (около 90), родовые и видовые дескрипторы для предикатов (больше 100) и семантические роли (около 50). Для первой очереди семантической разметки релевантные совокупности дескрипторов приписаны 3000 лексем. Разработан высокоэффективный и удобный в использовании комплекс программ для аннотирования текстов в полуавтоматическом режиме: каждая фраза сначала пропускается через

анализирующий модуль лингвистического процессора ЭТАП-3, а затем полученная синтаксическая структура редактируется лингвистом в специальной компьютерной среде. В самое ближайшее время доступ к глубоко аннотированному корпусу будет открыт в Интернете на общедоступном сайте.

4. Разработка интерактивной системы разрешения лексической неоднозначности для машинного перевода и других приложений. Основной проблемой при анализе текста на естественном языке является высокий уровень неоднозначности лексических единиц. Разработан новый модуль лингвистического процессора ЭТАП-3, позволяющий разрешить такую неоднозначность в режиме динамического взаимодействия человека и компьютера. В тех точках алгоритма анализа русского текста, где процессору недостает информации для определения того из потенциально возможных значений некоторого слова, в котором оно употреблено в этом тексте, активируется модуль интерактивного разрешения лексической неоднозначности, предлагающий пользователю системы на выбор несколько вариантов. Пользователь, руководствуясь своим знанием языка, а также экстралингвистической информацией и интуицией, выбирает нужный вариант. Получив от пользователя ответ, процессор продолжает обработку текущего предложения. Как показала широкая серия экспериментов с участием разработанного модуля, реализованный метод позволяет обеспечить высокую степень разрешения неоднозначности, которая не может быть достигнута чисто автоматическими средствами.

5. Разработка модуля синтактико-семантического анализа для системы многоязычной коммуникации в Интернете, основанной на Универсальном сетевом языке (UNL). Система многоязычной коммуникации в Интернете разрабатывается в рамках большого международного консорциума, состоящего из исследовательских групп из 14 стран. Система основана на искусственном языке для представления значений (UNL), выполняющем роль языка-посредника для всех рабочих языков системы. В ходе исследования был разработан массив правил, предназначенных для преобразования текстов на русском языке в гиперграфы UNL. Эти правила составляют важную часть UNL-модуля, разрабатываемого в качестве отдельной подсистемы лингвистического процессора ЭТАП-3. Вместе с русско-UNL словарем эти правила обеспечивают эффективное построение текстов UNL и в конечном счете организацию интерфейса между русским языком и всеми остальными рабочими языками системы. Совместно с партнерами из других стран были проведены эксперименты по гармонизации правил синтактико-семантического анализа для разных языков, и начата разработка единого корпуса текстов UNL, на основе которого можно будет создавать UNL-модули для новых языков.

РАБОТА С НАУЧНОЙ МОЛОДЕЖЬЮ

Сотрудники лаборатории в 2004 году вели преподавательскую работу в МГУ и Московском институте иностранных языков (акад. Ю.Д. Апресян), Российском государственном гуманитарном университете (Л.Л. Иомдин) и Мадридском политехническом университете (И.М. Богуславский). Ю.Д. Апресян и Л.Л. Иомдин руководили работами аспирантов по тематике лаборатории.

НАГРАДЫ

- Ю.Д. Апресяну присуждена Золотая медаль имени В.И. Даля за «Новый объяснительный словарь синонимов русского языка» в 3-х томах.

Институт проблем передачи информации РАН

ГРАНТЫ:

- **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".
- **Российский гуманитарный научный фонд (№ 04-04-00263):** "Нерешенные проблемы русского синтаксиса".
- **Российский фонд фундаментальных исследований (№ 04-06-80148):** "Разработка модуля синтактико-семантического анализа для системы многоязычной коммуникации в Интернете, основанной на Универсальном сетевом языке".
- **Российский фонд фундаментальных исследований (№ 04-07-90179):** "Разработка корпуса русских текстов, снабженного морфо-синтаксической разметкой".
- **Госконтракт по программе «Филология и информатика» (№ 10002-251/ОИФН-02/241-096/280504-384):** "Морфо-синтаксическая и лексико-семантическая разметка подмассива сбалансированного корпуса".
- **Контракт Минпромнауки России (№ 31):** "Модели и алгоритмы информационного взаимодействия в области генетики, лингвистики и цветного зрения".

ПУБЛИКАЦИИ В 2004 г.

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. 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.
4. Апресян Ю.Д. О семантической непустоте и мотивированности глагольных лексических функций // ВЯ, «Наука», 2004, № 4, 3-18.
5. 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.
6. Апресян Ю.Д. Интерпретационные глаголы: семантическая структура и свойства // Русский язык в научном освещении, «Языки славянской культуры», 2004, № 7, 5-22.
7. 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.
8. Иомдин Л.Л. Лексикографический портрет наречия что-то // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог'2003. М.: Наука, 2004, 246-251.

9. Иомдин Л.Л. Уроки машинного перевода для детей и взрослых. // Лингвистика для всех. Зимняя лингвистическая школа - 2004. М., НИИРО, 2004. 56-68.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.

Работы, находящиеся в печати

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.