

**НАУЧНЫЙ ОТЧЕТ ЗА 2007 ГОД**  
**по гранту Президента Российской Федерации**  
**для государственной поддержки**  
**молодых российских учёных - кандидатов наук (докторов наук)**  
**и их руководителей**  
**МК-9701.2006.6**  
**за счёт средств федерального бюджета**

**1. Номер гранта:**

МК-9701.2006.6

**2. Фамилия, имя, отчество:**

Митрофанова Ольга Александровна

**3. Объявленная ранее тематика:**

Проблемы количественной оценки семантической информации в словаре и в тексте

**4. Полученные за отчетный период научные (научно-технические) результаты:**

1. Совершенствование инструментов лингвистических исследований и развитие методов автоматической обработки языковых данных стимулирует решение задач, связанных с извлечением семантической информации из естественных языковых текстов. Одной из таких задач является осуществление автоматической классификации лексики (далее АКЛ) – процедуры, результаты которой востребованы во многих областях знаний о языке. АКЛ предоставляет лингвистам возможность использовать объективные данные об иерархической структуре лексикона, собранные при анализе представительных корпусов, и строить на основе этих данных формальные онтологии и лексикографические модули, применимые в процедурах автоматической обработки текстов и допускающие пополнение из корпусов. Использование инструментов АКЛ представляет интерес и в другом отношении: результаты кластеризации лексики позволяют решать вопросы автоматического индексирования текстов, тематического упорядочения документов в корпусах, способствует повышению качества информационного поиска в больших массивах текстов и пр.

В ходе реализации проекта исследованы теоретические и практические аспекты извлечения семантической информации из текстов, сформирована методологическая база для АКЛ, подтверждена целесообразность применения русскоязычного инструмента АКЛ в разных сферах фундаментальных и прикладных лингвистических исследований.

2. Проект направлен на построение русскоязычного ресурса АКЛ, который позволяет качественно выделять лексико-семантические классы слов из текстов разных объемов и разных типов, допускал бы классификацию лексики с различными условиями, открывал бы возможность использования результатов классификации в других системах автоматической обработки текста. Реализация проекта проводилась в несколько этапов: первым из них является создание инструмента АКЛ для работы с неразмеченными

текстами, далее производится усовершенствование инструмента АКЛ для обработки размеченных текстов и корпусов параллельных текстов.

К концу отчетного периода получены результаты обработки неразмеченных текстов с помощью разработанного инструмента АКЛ и ведется совершенствование данного ресурса.

3. Поставленная цель предполагала компьютерную реализацию алгоритма кластерного анализа. В нынешней версии ресурса АКЛ задействованы иерархический (агломеративный) метод кластеризации и неиерархический метод (К-средних). Выделение кластеров лексем в тексте производится на основе процедуры латентного семантического анализа (ЛСА). С лингвистической точки зрения, суть ЛСА заключается в возможности определения содержательной близости лексических единиц при сопоставлении их синтагматических свойств (иначе говоря, их сочетаемости с другими элементами контекста, дистрибуции). С инженерной точки зрения, ЛСА предполагает представление множества контекстов употребления исследуемых лексем как точек или векторов дистрибуций в N-мерном пространстве. Вычислив расстояния  $d$  между точками или сравнив вектора дистрибуций, можно получить количественную оценку тесноты семантических связей слов. При вычислении расстояний применяются различные меры близости. Ресурс допускает вычисление значений меры Евклида, меры Хэмминга, особое внимание также уделяется вычислению значения косинуса угла между векторами дистрибуций (поскольку в экспериментах данная мера зарекомендовала себя как наиболее надежная). Результаты измерений, сохраняемые в матрице расстояний, используются при кластеризации: чем ближе синтагматические свойства лексем (а стало быть, чем ближе их значения), тем меньше расстояние между векторами их дистрибуций и тем больше вероятность их объединения в один кластер. Сформированные таким образом кластеры лексем допускают дальнейшую лингвистическую интерпретацию.

4. В ходе подготовки экспериментов по АКЛ велась разработка соответствующего программного обеспечения. Программа АКЛ, созданная на языке Python, предусматривает три блока: блок предварительной обработки текста и вычисления расстояний между исследуемыми лексемами, блок иерархического кластерного анализа и блок кластерного анализа методом К-средних. Первый блок программы обеспечивает обработку входного текста. Прежде всего, обнаруживаются все вхождения исследуемых лексем в текст, затем производится автоматическое выделение границ контекстов в соответствии с заданной пользователем шириной контекстного окна, а также (по желанию пользователя) автоматическое определение весов элементов контекста. В дальнейшем для каждой лексемы  $l$  формируется множество контекстов ее употребления, которое представляется в виде вектора дистрибуции в N-мерном пространстве, где значения координат вектора соответствуют коэффициенту взаимной встречаемости  $l$  и других лексем из ее контекстов. Затем производится операция сравнения векторов дистрибуций всех исследуемых лексем. Итог работы программы на данном этапе – матрица расстояний между векторами дистрибуций для каждой пары лексем.

Второй и третий блоки программы отвечают за процедуры кластеризации иерархическим методом и методом



К-средних. При осуществлении иерархического кластерного анализа в тексте проводится пошаговое формирование совокупностей двух и более лексем, имеющих наиболее близкую дистрибуцию. Процедура повторяется до тех пор, пока все лексемы не объединятся последовательно в один кластер, или пока количество промежуточных кластеров (фактически, глубина иерархии) не достигнет числа, указанного пользователем. При кластеризации методом К-средних пользователь заранее указывает желаемое число кластеров. На первом шаге процедуры элементы кластеров выделяются случайным образом, затем вычисляются центры кластеров. При последующих шагах происходит перегруппировка кластеризуемых объектов относительно центров промежуточных кластеров. Процедура завершается, когда элементы займут стабильное положение в кластерной структуре.

Результаты кластеризации выводятся в виде многоуровневого списка слов с помощью скобочной записи. Наряду с этим пользователь получает данные о частотности исследуемых лексем в обрабатываемом тексте и значения расстояний во всевозможных парах лексем из анализируемого набора.

5. Для оценки эффективности работы и определения исследовательских возможностей разрабатываемого инструмента АКЛ была проведена серия экспериментов по извлечению и обработке данных из текстов разных типов:

- автоматическая классификация терминов-дескрипторов в научных текстах (на материале статей из русскоязычного неразмеченного корпуса по корпусной лингвистике);
- автоматическая классификация глагольной лексики в экспериментальном корпусе (на материале базовых глаголов русского языка и корпуса глагольных контекстов в двух версиях, размеченной и неразмеченной);
- автоматическая классификация лексики в параллельных неразмеченных текстах (на материале тематической группы существительных – обозначений живых существ, используемых в тексте оригинала и перевода повести-притчи Дж. Оруэлла "Скотный двор").

Проведенные эксперименты подтверждают эффективность разрабатываемого инструмента АКЛ.

6. Существующая версия ресурса АКЛ используется

- в проекте "Создание корпуса русскоязычных текстов по корпусной лингвистике" для определения логической структуры и систематизации терминологии предметной области, для классификации текстов по корпусной лингвистике и разработки формальной онтологии данной предметной области (проект реализуется совместно с ИЛИ РАН, Санкт-Петербург);
- в проекте "Снятие лексической неоднозначности в корпусах текстов русского языка" для автоматической классификации контекстов употребления предметных существительных в разных значениях (проект реализуется совместно с ВИНТИ РАН, Москва);
- в проекте "Автоматическая обработка корпусов параллельных текстов" для исследования семантической и стилистической близости текстов оригиналов и переводов на английский язык романов В.О. Пелевина (совместно с кафедрой прикладной лингвистики РГПУ им. А.И.Герцена, Санкт-Петербург)

7. В дальнейшем планируется:

- работа по техническому совершенствованию инструмента: расширение круга используемых методов кластеризации (например, MajorClust, Clustering by Committees), реализация алгоритма машинного обучения методом обратного распространения ошибки (backpropagation), добавление метрик при измерении расстояний между лексемами (введение меры Жаккара, коэффициента Дайса и пр.); модернизация пользовательского интерфейса (оптимизация системы запросов и добавление режима визуализации результатов кластеризации); введение модуля для автоматической классификации документов и пр.
- проведение лингвистических экспериментов с усложненными параметрами (АКЛ с учетом сочетаемостных предпочтений лексем, обработка неразмеченных и размеченных корпусов текстов большого объема, обработка неразмеченных и размеченных корпусов текстов различной жанровой и тематической принадлежности, построение формальных онтологий предметных областей на основе корпусов текстов, автоматическая классификация документов), дальнейшее использование инструмента АКЛ в решении практических задач по извлечению семантической информации из корпусов текстов.

## 5. Публикации грантополучателя за отчётный период по заявленной тематике:

- Общее количество публикации: 9
- монографий: 0
- учебников, учебных пособий: 0
- статей: 6
- тезисов докладов: 3
- количество публикаций в зарубежных научных изданиях: 6
- количество публикаций в научных изданиях стран СНГ: 2

| № п/п | Авторы, название публикации   | Вид публикации | Город, издательство   | Год издания | Кол-во страниц |
|-------|---|----------------|---|-------------|----------------|
| 1     | Митрофанова О.А.,<br>Мухин А.С.,<br>Паничева П.В.<br>Автоматическая классификация лексики в русскоязычных текстах на основе латентного семантического анализа | Статья         | Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции Диалог-2007. – М.: Издательство РГГУ, 2007. – С. 413–421. | 2007        | 9              |



|   |   |                   |   |      |   |
|---|---|-------------------|---|------|---|
| 2 | Митрофанова О.А.,<br>Белик В.В.,<br>Кадина В.В.,<br>Морозенко Ю.Я.<br>Исследование сочетаемости<br>предпочтений частотных лексем<br>русского языка на основе корпусных<br>данных      | Статья            | XXXVI Международная<br>филологическая конференция. Секция<br>математической лингвистики.<br>Издательство СПбГУ (в печати)   | 2007 | 6 |
| 3 | Mitrofanova O.,<br>Mukhin A.,<br>Panicheva P.,<br>Savitsky V.<br>Automatic Word Clustering in Russian<br>Texts  | Статья            | Eds. V. Matousek, P. Mautner et al. Text,<br>Speech and Dialogue. Proceedings of the<br>10th International Conference TSD 2007,<br>Plzen, Czech Republic, September 2007.<br>LNAI 4629. – Springer - Verlag, 2007. –<br>P. 85–91. | 2007 | 7 |
| 4 | Митрофанова О.А.,<br>Мухин А.С.,<br>Паничева П.В.,<br>Савицкий В.С.<br>Разработка лингвистического ресурса<br>для автоматической классификация<br>лексики (технология и эксперименты) | Тезисы<br>доклада | MegaLing–2006: Горизонты<br>прикладной лингвистики и<br>лингвистических технологий. Доклады<br>международной конференции. –<br>Симферополь: Издательство ДиАйПи,<br>2007. С. 137–140.   | 2007 | 4 |
| 5 | Митрофанова О.А.,<br>Белик В.В.,<br>Кадина В.В.,<br>Морозенко Ю.Я.<br>Сочетаемость предпочтения<br>частотных лексем русского языка по<br>корпусным данным                             | Тезисы<br>доклада | MegaLing–2007: Горизонты<br>прикладной лингвистики и<br>лингвистических технологий. Доклады<br>международной конференции. –<br>Симферополь: Издательство ДиАйПи,<br>2007. С. 141–143.   | 2007 | 3 |
| 6 | Виноградова Н.В.,<br>Митрофанова О.А.,<br>Паничева П.В.<br>Автоматическая классификация<br>терминов в русскоязычном корпусе<br>текстов по корпусной лингвистике                       | Статья            | Труды 9ой Всероссийской научной<br>конференции Электронные<br>библиотеки: перспективные методы и<br>технологии, электронные коллекции<br>(RCDL–2007). – Переславль -<br>Залесский, Россия, 2007. Т. 2. – 6 с.                     | 2007 | 6 |

|   |  |        |   |      |    |
|---|--|--------|---|------|----|
| 7 | Mitrofanova O.,<br>Panicheva P.,<br>Savitsky V.<br>Automatic Word Clustering in Russian<br>Texts based on Latent Semantic Analysis | Статья | Computer Treatment of Slavic and East<br>European Languages: Proceedings of the<br>4th International Seminar NLP,<br>Computational Lexicography and<br>Terminology: SLOVKO 2007.<br>Bratislava, Slovakia, October 25 –27,<br>2007. Bratislava: Tribun, 2007. P.<br>165–175. | 2007 | 11 |
| 8 | Mitrofanova O.,<br>Belik V.,<br>Kadina V.<br>Corpus Analysis of Selectional<br>Preferences in Russian                              | Статья | Computer Treatment of Slavic and East<br>European Languages: Proceedings of the<br>4th International Seminar NLP,<br>Computational Lexicography and<br>Terminology: SLOVKO 2007.<br>Bratislava, Slovakia, October 25 –27,<br>2007. Bratislava: Tribun, 2007. P.<br>176–182. | 2007 | 7  |
| 9 | Alexandrov M.,<br>Blanco X.,<br>Zakharov V.,<br>Mitrofanova O.<br>Clustering and Nooj Applications                                 |        | Презентация доклада URL:<br><a href="http://seneca.uab.es/jsastre/ppt/alexandrov_blanco_zakharov_mitrofanova.pdf">http://seneca.uab.es/jsastre/ppt/alexandrov_blanco_zakharov_mitrofanova.pdf</a>   | 2007 | 24 |

#### 6. Участие грантополучателя в отчётном году в научных конференциях и совещаниях по тематике проводимых исследований:

1) Семинар по корпусной лингвистике, ИЛИ РАН, Санкт-Петербург, Россия (свыше 40 участников)

Доклад: Митрофанова О.А., Мухин А.С., Паничева П.В. Автоматическая классификация лексики в неразмеченных русскоязычных текстах

2) Международная филологическая конференция. Секция математической лингвистики. Санкт-Петербург, Россия, 13 – 18 марта 2007 (свыше 800 участников)

Доклад: Митрофанова О.А., Белик В.В., Кадина В.В., Морозенко Ю.Я. Исследование сочетаемости предпочтений частотных лексем русского языка на основе корпусных данных

3) Международная конференция "Диалог–2007": Компьютерная лингвистика и интеллектуальные технологии, Москва, Россия, 30 мая – 3 июня 2007 (свыше 120 участников)

Доклад: Митрофанова О.А., Мухин А.С., Паничева П.В. Автоматическая классификация лексики в русскоязычных текстах на основе латентного семантического анализа

4) 2007 NooJ Conference, Barcelona, Spain, June 7 – 9, 2007 (свыше 50 участников)

Доклад: Alexandrov M., Blanco X., Zakharov V., Mitrofanova O. Clustering and Nooj Applications

5) 10th International Conference "TSD 2007": Text, Speech and Dialogue. Plzen, Czech Republic

September 3 –7, 2007 (свыше 110 участников)

Доклад: Mitrofanova O., Mukhin A., Panicheva P., Savitsky V. Automatic Word Clustering in Russian

Texts

6) Международная конференция "MegaLing–2007": Горизонты прикладной лингвистики и лингвистических технологий. Партенит, Крым, Украина, 24 – 28 сентября 2007 (свыше 130 участников)

Доклад: Митрофанова О.А., Мухин А.С., Паничева П.В., Савицкий В.С. Разработка лингвистического ресурса для автоматической классификация лексики (технология и эксперименты)

Доклад: Митрофанова О.А., Белик В.В., Кадина В.В., Морозенко Ю.Я. Сочетаемость предпочтения частотных лексем русского языка по корпусным данным

7) Всероссийская научная конференция "RCDL–2007": Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Переславль-Залесский, Россия, 15 – 18 октября 2007 (свыше 100 участников)

Доклад: Виноградова Н.В., Митрофанова О.А., Паничева П.В. Автоматическая классификация терминов в русскоязычном корпусе текстов по корпусной лингвистике

8) 4th International Seminar "NLP, Computational Lexicography and Terminology" SLOVKO 2007.

Bratislava, Slovakia, October 25 –27, 2007 (свыше 60 участников)

Доклад: Mitrofanova O., Panicheva P., Savitsky V. Automatic Word Clustering in Russian Texts based on Latent Semantic Analysis

Доклад: Mitrofanova O., Belik V., Kadina V. Corpus Analysis of Selectional Preferences in Russian

**7. Выполнение исследований по ФЦП "Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса на 2007-2012 годы" и/или по другим ФЦП, академическим, отраслевым программам; по приоритетным направлениям; по грантам РФФИ и РГНФ, а также по международным грантам за отчетный период:**

Общее количество: 0

**8. Адреса ресурсов в Internet, подготовленных грантополучателем:**

<http://oa-mitrofanova.narod.ru/>

**9. Преподавательская деятельность грантополучателя:**

Должность: доцент кафедры математической лингвистики филологического факультета Санкт-Петербургского государственного университета

Лекционные курсы:

- 1) Формальная семантика;
- 2) Формальные методы в лингвистике;
- 3) Уровни лингвистического анализа.

Руководство исследовательской работой студентов:

всего – 21 проект, из них

5 – магистерские диссертации,


6 – дипломные проекты,

10 – курсовые проекты.

#### **10. Участие грантополучателя в экспедициях:**

0

**Грантополучатель**

 / Митрофанова О. А. /