ИНТЕРНЕТ/ИНТРАНЕТ СИСТЕМА ВЕДЕНИЯ И ОБРАБОТКИ КОРПУСА ТЕКСТОВ С РАЗВИТЫМИ ЛИНГВИСТИЧЕСКИМИ СРЕДСТВАМИ НА ОСНОВЕ J2EE И ORACLE9i ТЕХНОЛОГИЙ

С.М. Козиенко[1], С.А. Яблонский [1,2]

([1]Петербургский государственный университет путей сообщения,

[2]ЗАО "Руссикон")

Создание систем для ведения и обработки корпуса текстов является достаточно традиционной задачей корпусной лингвистики для романских и германских языков. За последнее десятилетие в сети Интернет и на CD появились большие электронные коллекции различных текстов на русском языке: художественных, научных, правовых и других. Большинство электронных библиотек отличает крайняя бедность средств работы с опубликованными текстами, что не дает в полной мере использовать их как для научной работы, так и в целях образования, особенно филологического. Кроме того, те немногие электронные библиотеки, в которых предусмотрены средства работы с текстом, моментально становятся предметом продажи, что также не позволяет широко использовать их в сети Интернет для указанных выше целей.

Задача рассматриваемой системы заключается в разработке средств создания, ведения и обработки простых и аннотированных моно- и многоязычных корпусов текстов большой размерности (до сотен миллионов слов) при помощи развитых средств линвистической поддержки и поиска с возможностью представления результатов в сети Интернет/Интранет.

В более широком смысле система может рассматриваться как основа любой поисковой системы в Интернете, а также системы документооборота.

Хранение документов в базе данных Oracle 9i

В настоящее время при построение корпусов текстов широко используется XML формат представления документов (например, XML Corpus Encoding Standard, http://www.xml-ces.org, [8]), хотя по-прежнему существует огромное количество текстов, создаваемых в таких рапространеных форматах как HTML, RTF, DOC, PDF и др.

Существуют несколько стратегий хранения корпусов текстов (документов) в базе данных Oracle [2,7]:

1. Хранение документов как отдельных неделимых объектов (**документоцентричный** подход):

   - документы хранятся вне базы данных, а в базе данных строится только система индексации и поиска;
   - документы хранятся как данные типа CLOB, BLOB или XMLtype.

   Хранение документов в базе данных как неделимых объектов подходит в том случае, когда их содержание статично и, что существенно, любое обновление документа сводится к его перезаписи в базе данных. Типичные примеры таких

документов - статьи, книги, технические руководства, контракты и т.д. Это документы в обычном значении этого слова, они хранятся в базе данных целиком и поставляются из нее вовне также целиком. Oracle может не только хранить документы различных форматов (до 150), но и организовывать по ним эффективный поиск, в том числе - с использованием морфологии языка.

2. Хранение элементов документов как данных (то есть собственно данных, без тэгов разметки того или иного формата) в объектно-реляционном или реляционном представлении, фактически, в реляционных таблицах базы данных (**датацентрический** подход).

   Если документ структурно корректен и содержит элементы, которые могут обновляться и вообще использоваться по отдельности, а не как единое целое, то такой документ можно назвать датацентрическим. Обычно такие документы включают один или несколько элементов со сложной структурой. Примерами могут послужить бланки заказов, финансовые счета и т.д., то есть документы на базе сложных форм.

3. Смешанное хранение документов и данных.

   Наконец, если необходимо обрабатывать документы смешанных типов, когда имеются как структурированные, так и неструктурированные данные, рассматриваемые, тем не менее, как единый документ, можно использовать представления (view) Oracle. Они позволяют конструировать объекты “на лету”, комбинируя данные, которые хранятся в различном виде.

Документоцентрический подход наиболее полно соответствует задаче создания и ведения корпуса текстов. Он и был использован в настоящей работе. Ниже подробно рассмотрен вариант реализации корпуса, когда тексты (документы) хранятся вне базы данных, а в базе данных строится только система индексации и поиска.

Описание системы

На рис. 1 показана общая архитектура системы.

Помимо сервера баз данных Oracle9i используется Web-сервер, который принимает, обрабатывает запросы и формирует ответы на них. Web-сервер может располагаться как на отдельном компьютере, так и на сервере БД Oracle.

Система, состоит из следующих подсистем:

- администратора системы,
- индексирования и поиска документов,
- тезауруса/рубрикатора,
- морфологического анализатора и лемматизатора,
- обработки новых слов,
- конкорданса и толкового словаря,
- подсистемы интерфейса.

Рассмотрим некоторые из них.

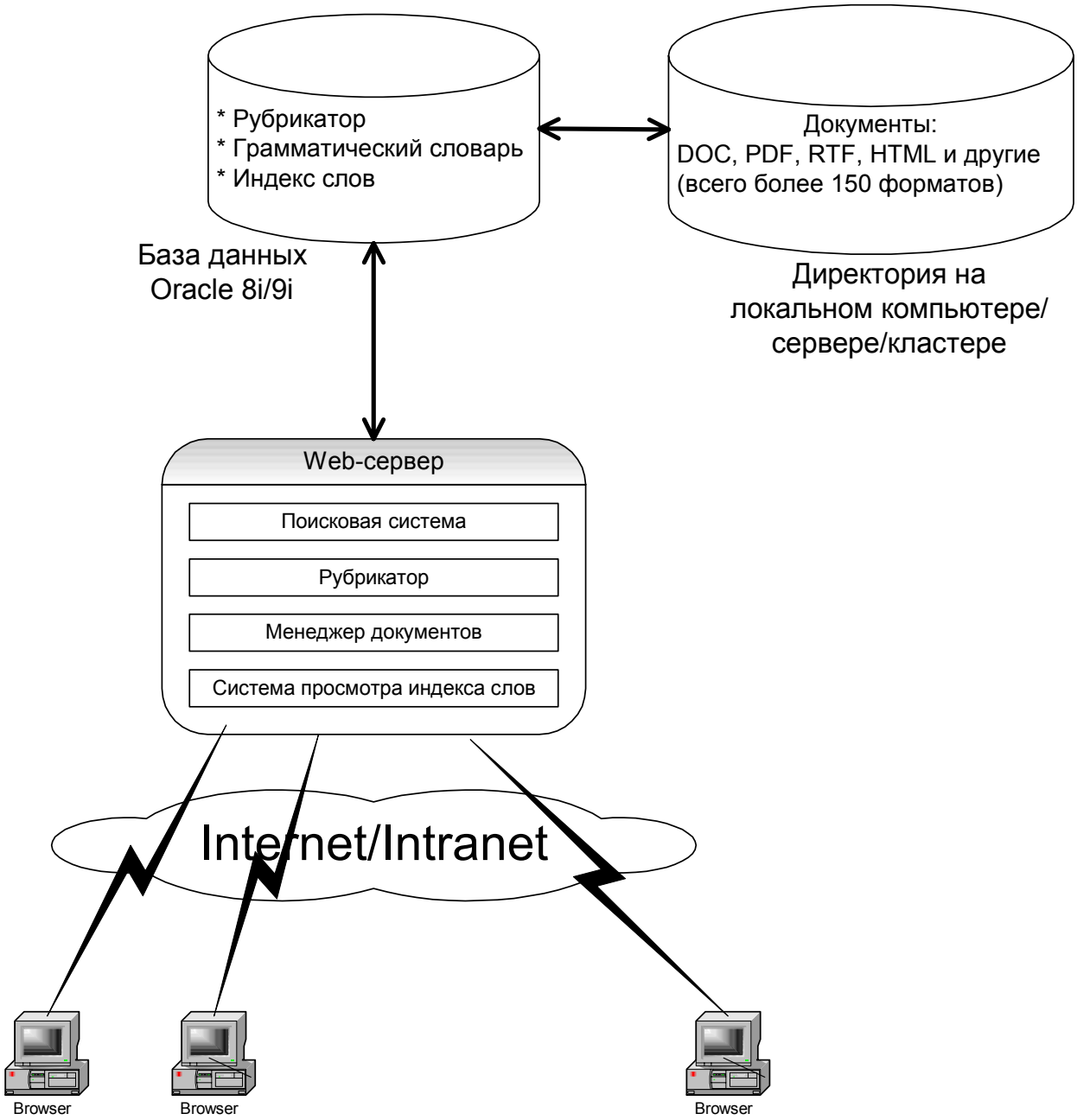

Рис. 1. Общая архитектура системы

Администрирование системы, поиск, индексирование документов

На рис. 2 показана диаграмма состояний для пользовательского интерфейса подсистемы поиска. Под страницами подразумеваются HTML-страницы соответствующих подсистем. Диаграммы иллюстрируют и основные функции рассматриваемых подсистем.

Поисковая подсистема (рис.3) позволяет искать документы (тексты):

- по словам, объединенным логическими связками и скобками;
- по фразе, заключенной в кавычке;
- по времени добавления документа;
- по языку, на котором написан документ;
- по форматам документов.

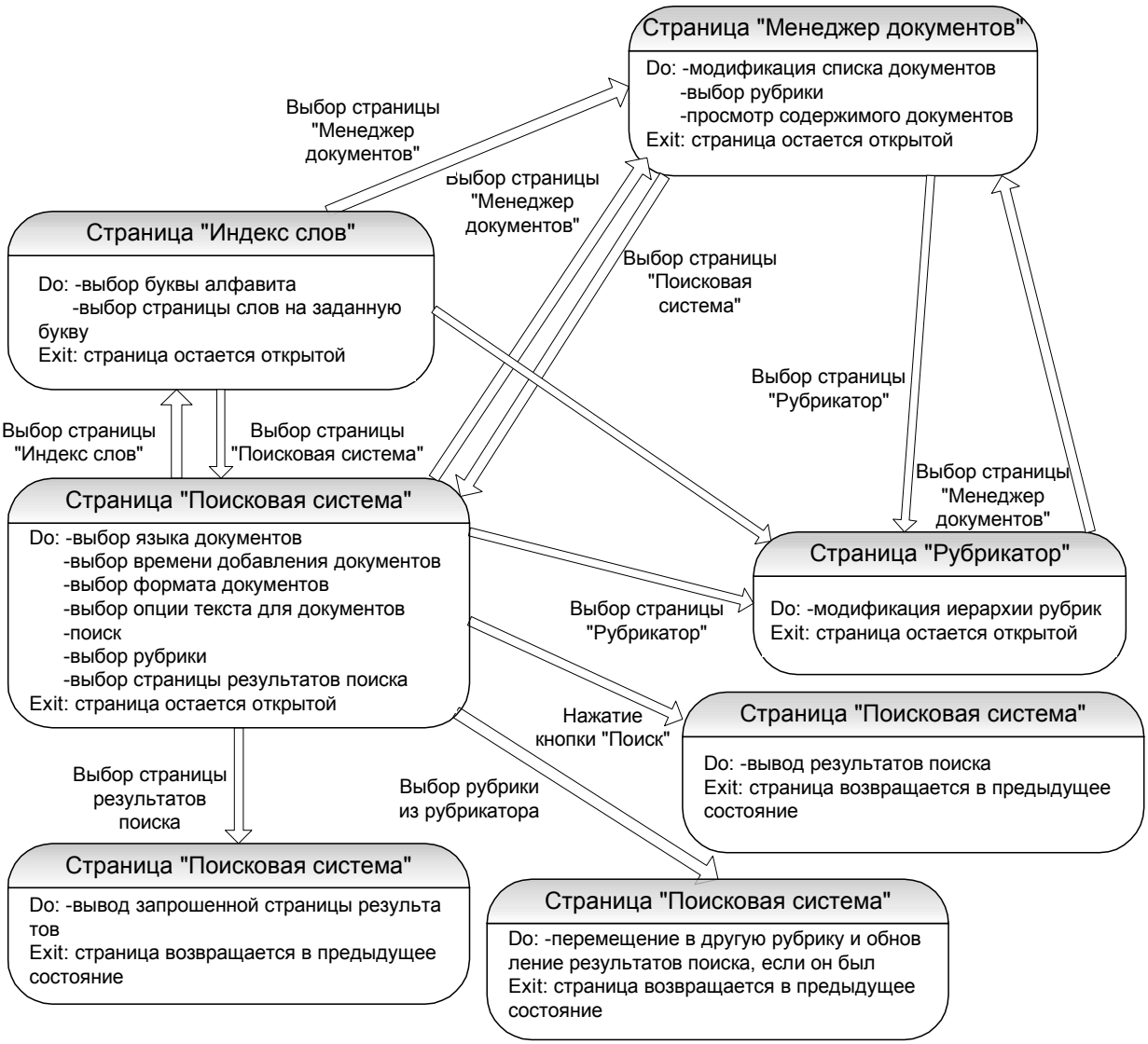

Рис. 2

Морфологический анализатор и нормализатор

При морфологическом анализе русскоязычного текста всем словам анализируемого текста сопоставляются леммы (нормализованные словоформы) с соответствующей грамматической информацией (род, число, падеж и т.п.) [5,8]. Размер используемого грамматического словаря - 150 тысяч лемм [5], что обеспечивает более чем 95-99 % покрытия правовых и СМИ текстов, и 92-94 % - художественных текстов. Для новых слов в автоматизированном режиме порождаются гипотезы, содержащие правильную лемму и парадигму.

Включение лемм в индекс вместо словоформ позволяет резко сократить размер индекса при больших объемах хранимых текстов. При построении индекса система просматривает все тексты корпуса и преобразует каждую словоформу в соответствующую ему нормальную форму - лемму (для существительных это - именительный падеж единственного числа, для глаголов - неопределенная форма и т.д.), на основании чего и формируется индекс. С другой стороны, при поиске по одной заданной словоформе находятся все документы, содержащие любые возможные словоформы из парадигмы, включающей заданную словоформу.

Рис. 3. Поисковая система образовательного портала

Конкорданс и толковый словарь

Подсистема поиска, по сути, реализует конкорданс тестов, размещенных в системе. Конкорданс – это список словоформ/лемм, встречающихся в тексте, расположенных в алфавитном порядке, где слово даётся с его словесным окружением (полное предложение или его часть). Совместно с толковым и грамматическим словарем конкорданс позволяет полнее описать слово, что особенно важно в обучающих системах и для специалистов-филологов.

Разработка UML-диаграмм модели данных

Проектирование системы велось на основе объектно-ориентированного подхода с использованием CASE-средств разработки Rational Rose 2001 Enterprise компании Rational Software Corp.[3]

Каждый документ представляется в базе данных в виде следующих основных полей:

- Автор: ФИО;
- Название документа: название/место/дата издания;

- Имя файла и путь к нему: путь к документу на сервере;
- Рубрика из рубрикатора-тезауруса;
- Язык, на котором написан документ;
- Дата добавления документа.

На рис. 6 представлен фрагмент UML-диаграммы модели данных поисковой системы.

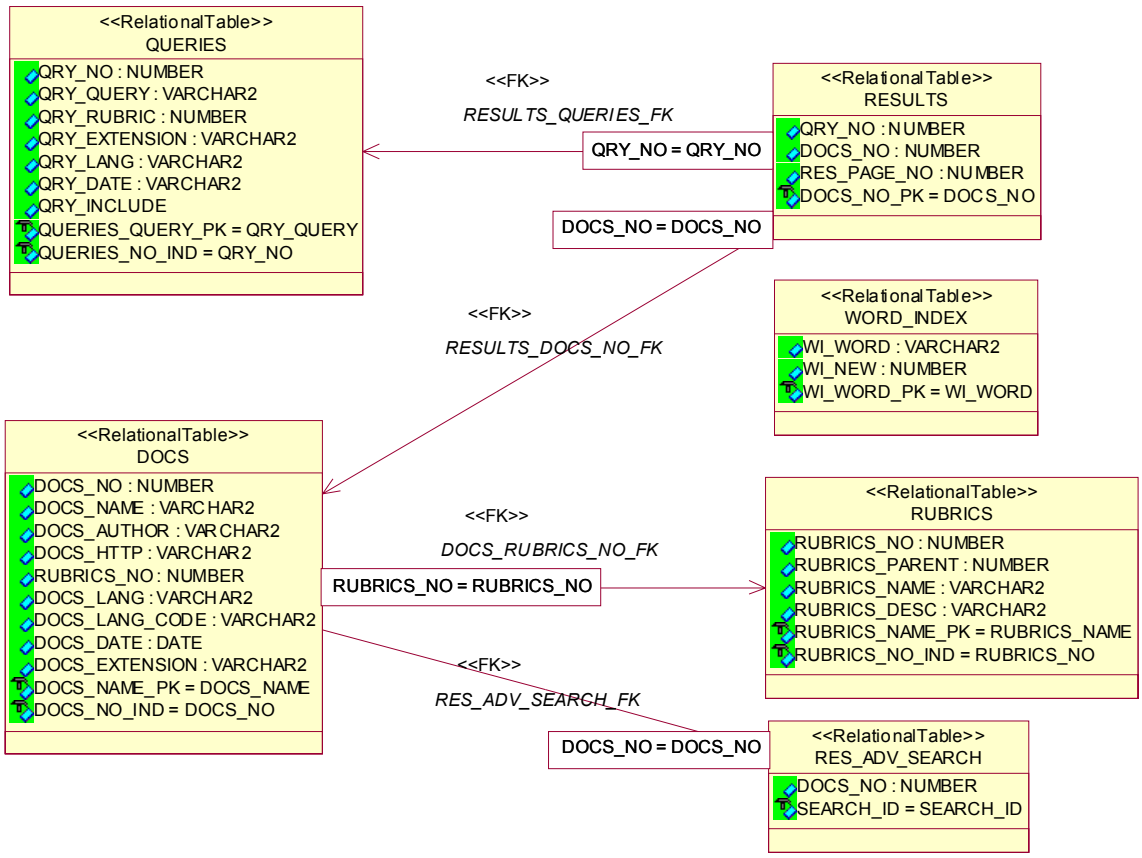

Рис. 6. Упрощенная UML-диаграмма модели данных поисковой системы

Описание таблиц, используемых поисковой системой:

- Таблица **DOCS —** содержит информацию об индексируемых документах

Обозначение в модели	**Сущность**	**Описание**
DOCS_NAME	Наименование документа	Содержит наименование статьи или документа
DOCS_AUTHOR	Автор	Содержит имя автора, который составил документ
DOCS_HTTP	Путь	Содержит путь на сервере к документу
DOCS_LANG	Язык	Содержит краткое наименование языка, на котором составлен документ
DOCS_LANG_CODE	Кодировка	Содержит наименование кодировки документа
DOCS_DATE	Дата добавления	Содержит дату добавления документа
DOCS_EXTENSION	Расширение	Содержит расширение документа

	файла	
RUBRICS_NO	Рубрика	Содержит ссылку на рубрику в таблице RUBRICS, к которой принадлежит документ

- Таблица **QUERIES –** содержит тексты и параметры запросов. Эта таблица является кэшем для запросов, что является оправданным, если в качестве результата будет возвращено более 10 страниц или если несколько пользователей производят одинаковые запросы.

Обозначение в модели	Сущность	Описание
QRY_QUERY	Текст запроса	Содержит текст запроса
QRY_RUBRIC	Рубрика	Содержит ссылку на рубрику в таблице RUBRICS, в которой осуществляется поиск
QRY_EXTENSION	Расширение файла	Содержит расширение файла, по которому будет осуществляться поиск
QRY_LANG	Язык	Содержит язык документов, которые будут включены в результат поиска
QRY_DATE	Дата добавления	Содержит дату, начиная с которой документы будут включены в результат поиска

- Таблица **RESULTS –** содержит результаты поиска.

Обозначение в модели	Сущность	Описание
QRY_NO	Текст запроса	Ссылка на текст запроса в таблице QUERIES
DOCS_NO	Документ	Ссылка на документ в таблице DOCS
RES_PAGE_NO	Номер страницы	Содержит номер страницы, которой принадлежит данное решение

- Таблица **RUBRICS –** содержит информацию для рубрикатора.

Обозначение в модели	Сущность	Описание
RUBRICS_PARENT	Рубрика-родитель	Ссылка на рубрику в таблице RUBRICS, для которой данная рубрика является подрубрикой
RUBRICS_NAME	Рубрика	Содержит наименование рубрики
RUBRICS_DESC	Описание	Содержит краткое описание рубрики

- Таблица **WORD_INDEX –** содержит полный список слов, которые содержатся в документах

Обозначение в модели	Сущность	Описание
WI_WORD	Слово	Содержит слово
WI_NEW	Флаг	Содержит флаг, который указывает, есть ли данное слово в словаре русского языка

- Таблица **RES_ADV_SEARCH –** временно содержит промежуточные результаты для улучшенного поиска

Обозначение в модели	Сущность	Описание
SEARCH_ID	Идентификатор поиска	Содержит идентификатор поиска для определения результатов
DOCS_NO	Документ	Ссылка на документ в таблице DOCS

Заключение

Специфика системы заключается в следующем:

- корпус текстов достаточно полно представляет широкий спектр художественных технических, правовых и СМИ документов на русском языке [6];
- используются платформно-независимые технологии J2EE (Servlets и Java Server Pages) и Oracle9i [2];
- применяется объектно-ориентированный подход при проектировании системы на основе UML-нотаций [1,3,7];
- используется встроенное в СУБД Oracle 9i средство работы с текстовыми данными Oracle Text [2], что позволяет применять:
    - встроенные лингвистические средства работы с текстами ряда европейских языков (распознование числа, времени и других типов слов, а также грамматических форм, неверно написанных слов и слов, сходных по звучанию);
    - 150 конвертеров наиболее широко распространенных форматов текстовых файлов (ASCII, doc, rtf, pdf и пр.);
    - кроме точного поиска по слову или словосочетанию с выполнением известных булевых операций, также поиск с упорядочиванием по релевантности и заданием списков стоп-слов;
    - встроенные и создаваемые пользователями простейшие типы тезаурусов для поиска синонимов и тематически близких слов;
    - анализ содержания документа корпуса и автоматическое выделение его ключевых тем с созданием тематического резюме.
- используются апробированные лингвистические ресурсы и программы ЗАО “Руссикон”[5], что позволяет использовать средства OracleText и для русскоязычных текстов (поскольку сам OracleText не поддерживает русский язык);
- реализован полнотекстовый поиск для русскоязычных текстов с учетом морфологии русского языка;
- использование нормализатора позволяет создавать компактный индекс для русскоязычных текстов за счет хранения лемм (а не словоформ).

Применяемые решения позволяют работать как с корпусом русских текстов, так и многоязычными корпусами тестов за счет использования Unicode (очевидно, что для каждого языка должен быть включен свой набор лингвистических ресурсов).

Список литературы

1. Booch, G., Rumbaugh, J., and Jacobson, I., 1998. The Unified Modeling Language user guide, Addison-Wesley.
2. Oracle9i Database Documentation (Release 9.0.2), 2002.
3. Rational Rose Enterprise Edition 2001, Documentation.
4. Sullivan D., 2001. Document Warehousing and Text Mining, John Wiley & Sons, 542 p.

5. Yablonsky S.A., 1998. Russicon Slavonic Language Resources and Software. In: A. Rubio, N. Gallardo, R. Castro & A. Tejada (eds.) *Proceedings First International Conference on Language Resources & Evaluation,* Granada, Spain.
6. Yablonsky S.A., 2000. Russian Monitor Corpora: Composition, Linguistic Encoding and Internet Publication. Proceedings Second International Conference on Language Resources & Evaluation, Athens, Greece, 2000.
7. Yablonsky S.A., 2002. Corpora as Object-Oriented System. From UML-notation to Implementation. In: *Proceedings Third International Conference on Language Resources & Evaluation,* Las Palmas, Spain.
8. Ide, N., Romary L., 2001. XML Support for Annotated Language Resources. In: *Linguistic Exploration,* Workshop on Web-Based Language Documentation and Description, Dec 12 - Dec 15, 2000, University of Pennsylvania Philadelphia, Pennsylvania, USA.