Т.В. Батура, О.В. Корда, Ф.А. Мурзин, А.А. Позименко*

ИССЛЕДОВАТЕЛЬСКАЯ СИСТЕМА ДЛЯ АНАЛИЗА ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

ВВЕДЕНИЕ

Исследования в области автоматической обработки текста (АОТ) и формализации естественных языков, планомерно продвигаясь от самых простых методов анализа к более сложным, постепенно приближаются к такому уровню обработки текста, на котором уже возможно представление текста не просто в виде последовательности слов, а единым целым, обладающим неким смыслом, что уже соответствует человеческому восприятию. Увеличение вычислительных мощностей сделало возможным применение трудоёмких лингвистических алгоритмов на больших объемах данных.

Основной целью данной работы является разработка исследовательской системы, реализующей новые подходы к анализу текстов на естественном языке. В результате построено приложение «испытательный стенд», использующее существующие системы для «первичного» анализа текста и реализующее собственные подходы к анализу и представлению естественного языка.

1. СИСТЕМЫ ОБРАБОТКИ ТЕКСТОВ

В настоящее время лингвистами сформулированы различные теории, позволяющие в какой-то степени формализовать естественный язык. В основном, суть этих теорий сводится к тому, что предложению в тексте сопоставляются различные конечные объекты — графы или, в общем случае, конечные модели, которые, как принято считать [1], отражают смысл предложений.

* murzin@academ.org, tbatura@ngs.ru

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 05-08-33448).

1.1. Общие принципы систем обработки текстов

Компоненты, составляющие структуру систем анализа текстов — лингвистические процессоры, которые последовательно обрабатывают входной текст. Вход одного процессора является выходом другого [2].

Выделяются следующие компоненты:

- графематический анализ — выделение слов, цифровых комплексов, формул и т.д.;
- морфологический анализ — построение морфологической интерпретации слов входного текста;
- синтаксический анализ — построение дерева зависимостей всего предложения;
- семантический анализ — построение семантического графа текста.

Для каждого уровня разрабатывается свой язык представления. Язык представления обычно состоит из констант и правила их комбинирования. На графематическом уровне используются константы, называемые графематическими дескрипторами (ЛЕ — лексема, ЦК — цифровой комплекс и т.д.). На морфологическом уровне — граммемы (**рд** — родительный падеж, **мн** — множественное число). На синтаксическом — названия отношений (**subj** — отношение между подлежащим и сказуемым, **circ** — обстоятельство). О семантическом анализе будет сказано ниже.

Основой для построения уровней служат результаты работы предыдущих этапов, но, что важно, последующие анализаторы также могут улучшить представление предыдущих. Например, для какого-то предложения синтаксический анализатор не смог построить полного дерева зависимостей, тогда, возможно, семантический анализатор сможет спроектировать построенный им семантический граф на синтаксис.

Такой многоуровневый подход позволяет предложить критерии оценки систем машинного перевода. Вполне можно утверждать, что разработчики ФРАП [6, 7] показали, что для достижения адекватности перевода (равенство по смыслу входному тексту) и грамматической правильности выходной фразы необходимо присутствие всех пяти этапов, причем адекватность перевода можно гарантировать только после работы «глубоких» — синтаксического и информационного — анализаторов.

1.2. Система Диалинг

Система Диалинг разрабатывалась как система русско-английского перевода с 1999 по 2002гг. на базе ООО «Диалинг» [2, 3]. В разное время в

работе над системой принимали участие 22 специалиста, большинство из которых — известные учёные-лингвисты [4, 5].

Как и все современные системы обработки текста, Диалинг включает в себя основные этапы анализа текста: графематический, морфологический и синтаксический, а также ещё один, не так давно появившийся, семантический этап. В отличие от морфологического и синтаксического, на семантическом этапе появляется формальное представление смысла текста.

За основу системы автоматического русско-английского перевода Диалинг были взяты система французско-русского автоматического перевода (ФРАП), разработанная в ВЦП совместно с МГПИИЯ им. М. Тореза в 1976–1986гг. [6, 7], и система анализа политических текстов на русском языке — ПОЛИТекст, разработанная в центре информационных исследований в 1991–1997гг. [8].

Система ФРАП содержала полную цепочку анализа текста, вплоть до семантического, который был реализован только частично. В системе ФРАП был разработан и опробован семантический аппарат, на основе которого в системе Диалинг был создан оригинальный метод семантического анализа — метод полных вариантов. В центре семантического аппарата ФРАП находятся два перечня (вернее, две грамматики): семантических характеристик (СХ) и смысловых отношений (СО). Используется минимальное количество семантических характеристик: ВЕЩВО (“вещество”), ИЗМ (“изменение”), ИНТЕЛ («интеллектуальность»), ИНФ (“информация”) и т. д.; слова характеризуются по признаку принадлежности к одному или нескольким классам. СХ обеспечивают проверку семантического согласования при интерпретации связей в тексте.

Система ПОЛИТекст была направлена на анализ официальных документов на русском языке и содержала полную цепочку анализаторов текста: графематический (первичный анализ), морфологический, синтаксический и частично семантический. Из системы Диалинг был частично заимствован графематический анализ, но адаптированный под новые стандарты программирования. Программа морфологического анализа была написана заново, поскольку скорость работы была низкой, но сам морфологический аппарат не изменился.

2. ПОСТАНОВКА ЗАДАЧИ

Целью данной работы являлось создание исследовательской системы для анализа текстов на естественном языке. Под этим подразумевается не некото-

рый продукт, готовый к применению в прикладных задачах, а система, в которой реализовываются различные подходы для анализа текстов, т.е. целью проекта является получение наработок в области анализа текстов на естественном языке. Эти наработки могут быть использованы в других проектах, а также для дальнейших исследований структуры естественных языков.

Конкретной целью работы является программная система, реализующая следующие подходы к семантическому анализу текстов на естественном языке:

- сопоставление тексту набора предикатов узкого исчисления (лексических функций, грамматических предикатов и др.) [9–12];
- представление предложений с помощью деревьев с пометками на основе определения слова (словарной статьи из словаря Ожегова).

Система должна обеспечивать:

- загрузку текста;
- морфологический и синтаксический анализы текста;
- реализацию перечисленных подходов к анализу текста;
- графический вывод результатов анализа.

Отметим, что морфологический и синтаксический анализы производятся посредством использования внешних модулей (системы Диалинг). Они необходимы в системе формирования грамматических предикатов и для других целей.

3. ПРЕДСТАВЛЕНИЕ ПРЕДЛОЖЕНИЙ С ПОМОЩЬЮ ДЕРЕВЬЕВ С ПОМЕТКАМИ

Считаем, что поисковый запрос представляет собой совокупность предложений на естественном языке. Эту совокупность предложений можно расширить, используя словарные статьи из толкового словаря (например, словаря Ожегова), т.е. фактически приписать определения отдельных слов. Следующий этап — представление данных предложений в виде помеченных деревьев. Вершины помечаются словами, а ребра — вопросами, задаваемыми от одного слова к другому.

Далее имеется текст достаточно большого объема, из которого необходимо выбрать предложения по тематике поискового запроса и таким образом сформировать аннотацию или решить, является ли текст релевантным к данному запросу. Для этого предложения данного текста также могут быть представлены в виде деревьев (вообще говоря, необязательно все предло-

жения, а выборочно, по некоторым критериям). После этого необходимо сопоставление на похожесть (соответствие) деревьев из запроса и деревьев, возникших из текста.

Для аннотации выбираются предложения, которые соотносятся по теме, имеют похожие структуры и т.д. На основании подобных идей можно судить о релевантности.

Данный вопрос еще требует серьезной доработки. Часть материала представлена ниже.

3.1. Структура словарной статьи в словаре С.И. Ожегова

Для описания структуры словарных статей необходимы некоторые обозначения.

Наклонным шрифтом в скобках будем обозначать пояснение, некоторую характеристику между двумя словами или между словом и группой слов, относящейся к этому слову. Полужирным шрифтом записываем вопросы. Вопросы можно было бы записывать везде, но в этом нет необходимости. Вопросы указаны только в тех случаях, когда они несут большую информативность, нежели пояснения, или для того, чтобы провести общую линию в схемах статей (наличие ограниченного числа вопросов, которые часто повторяются). Полужирным и наклонным одновременно записаны слова, которые подразумеваются, но не присутствуют в толковании. Это удобно при схематической записи статьи, в которой встречаются причастия (причастия легко представить через глагол, так как они производные от глаголов).

Если вопрос задан к одной самостоятельной части речи, то пишем *синоним* или *синоним, но из другой части речи*. Если вопрос задан к словосочетанию или целому предложению, то этот фрагмент разбирается самостоятельно.

К однородным членам предложения и от однородных членов предложения задается каждый вопрос в отдельности. Например:

«употребляется для указания на расстояние или время, отделяющие одно пространство или событие от другого».

для указания -> (**на что?**) на расстояние

для указания -> (**на что?**) на время

Если между однородными членами предложения стоит союз, и слово относится к каждому из них, то ставится знак $\vee$. Например:

на расстояние $\vee$ время -> (*признак предмета*) отделяющие одно пространство или событие от другого.

В скобки […] заключена конструкция (она является целиком одним членом предложения), состоящая из набора слов, целиком относящаяся к одному слову, но не сложное предложение.

В скобки {…} заключена часть сложного предложения (не первая), т.е., если предложение сложное, в нем обязательно встречаются такие скобки.

Предложение или конструкцию записываем полностью после слова, от которого задается вопрос. Затем в скобках идет детальный разбор.

Если возможно задать два вопроса или более, то они записаны через «;».

Структурная запись (запись, как в языках программирования) помогает разделить толкование на семантические слои. В каждом таком слое обязательно присутствует центральное (главное) слово, т.е. слово, от которого задаётся вопрос.

Типичная словарная статья, как правило, может быть представлена в виде $\langle t, m, s \rangle$, где

t — заглавное слово в словарной статье;

m — слово, выполняющее служебную функцию;

s — толкование значения слова (определение слова).

Пример.

ДО. *предлог с род. п.*

1. Употребляется для указания на расстояние или время, отделяющие одно пространство или событие от другого.

Здесь t = «до», m = «употребляется», s — остальная часть предложения.

Слово, выполняющее служебную функцию, иногда отсутствует, но подразумевается. Например, в статье «**ЗАГРАНИЧНЫЙ**». Тогда в толковании дается сразу синоним или синоним, но из другой части речи.

ЗАГРАНИЧНЫЙ. Относящийся к зарубежным странам, зарубежный.

Здесь t = «заграничный», m отсутствует, s — остальная часть предложения.

Случай, когда слово, выполняющее служебную функцию, присутствует, более всего характерен для толкования предлогов. Гораздо реже подобная ситуация встречается в объяснении наречий. Большинство статей для наречий начинается со слов:

в места, в месте, в место, из мест, из места, на место, от места

в направлении, по направлению

в сторону, на стороне, со стороны, со сторон

в части

из источника

на пространство
на расстоянии, с расстояния
по линии
по поверхности.

Большая часть статей для существительных начинается со слов:
полоса (*чего-то*)
предмет (*чего-то*)
род (*чего-то*)
слой (*чего-то*)
часть (*чего-то*).

В толкованиях прилагательных характерно наличие отношения *признак предмета* (так как объяснение часто дается не через прилагательные, а через причастия), а глаголов, как и всех вышеперечисленных частей речи, — наличие отношения *синоним*.

Паре $\langle m, s \rangle$ может быть сопоставлена схема, аналогичная схеме синтаксического разбора. Если m отсутствует, то такая схема представляет собой просто синтаксический разбор предложения s.

ДО. *предлог с род. п.*

1. Употребляется для указания на расстояние или время, отделяющие одно пространство или событие от другого.

употребляется -> (*цель*, **для чего?**) для указания
для указания -> (**на что?**) на расстояние
∨
для указания -> (**на что?**) время
на расстояние ∨ время -> (*признак предмета*) отделяющие
одно пространство или событие от другого
[
(***которые***) отделяют
отделяют -> (**что?**) пространство
∨
отделяют -> (**что?**) событие
пространство ∨ событие -> (*признак предмета*) одно
пространство ∨ событие -> (**от чего?**) от другого
]

Или другой пример, в котором m отсутствует.

ПЕРЕНОСИЦА. Верхняя часть носа, примыкающая ко лбу и образующая углубление между лбом и носом.

переносица -> (*синоним*) часть носа
часть -> (*признак предмета*) верхняя
часть -> (*признак предмета*) примыкающая ко лбу и образующая углубление между лбом и носом
[
(***которая***) примыкает
примыкает -> (**к чему?**) ко лбу
&
(***которая***) образует
образует -> (**что?**) углубление
углубление -> (**между чем?**) между лбом
углубление -> (**между чем?**) между носом
]

Таким образом, каждый узел схемы представляет собой либо пару $\langle w, q \rangle$, где *w* — слово (оно стоит перед стрелкой), *q* — вопрос (стоит в скобках после стрелки), либо $\langle s \rangle$ — а) предложение, или б) причастный, деепричастный оборот, или в) несогласованное определение, или г) сравнительная, превосходная степени сравнения прилагательных или наречий (их сложная форма образования), или д) некоторые неразрывные словосочетания.

Предположим, что дано предложение или часть предложения $w_1 p_1 w_2 p_2 ... w_n p_n$, где w_i — слова, p_i — разделители, т.е. пробелы или знаки препинания. Тогда можно рассмотреть упорядоченный кортеж $\langle w_1, ..., w_n \rangle$.

Переходы вида $\begin{matrix} \langle w_i, q_i \rangle \\ \uparrow \\ \langle w_j, q_j \rangle \end{matrix}$ и $\begin{matrix} \langle w_i, q_i \rangle \\ \uparrow \\ \langle s_j \rangle \end{matrix}$ могут рассматриваться как правила формальной грамматики Хомского. На каждом шаге вопрос (нетерминальный символ грамматики) заменяем словом или предложением (терминальным символом). Предложения $\langle s_j \rangle$, которые подставляются в схеме вместо вопроса q_i, могут быть разобраны аналогичным образом.

3.2. Пример построения дерева

Рассмотрим пример со словом «междуречье».

Местность между двумя или несколькими реками, включающая водоразделы и прилегающие склоны долин.

междуречье -> (*синоним*) местность

местность -> (*признак предмета;* **между чем?**) между реками
между реками -> (*количество,* **сколькими?**) двумя
∨
между реками -> (*количество,* **сколькими?**) несколькими
местность -> (*признак предмета*) включающая водоразделы и прилегающие склоны долин
[
(***которая***) включает
включает -> (**что?**) водоразделы
&
включает -> (**что?**) склоны
склоны -> (*признак предмета*) прилегающие
склоны -> (**чего?**) долин
]

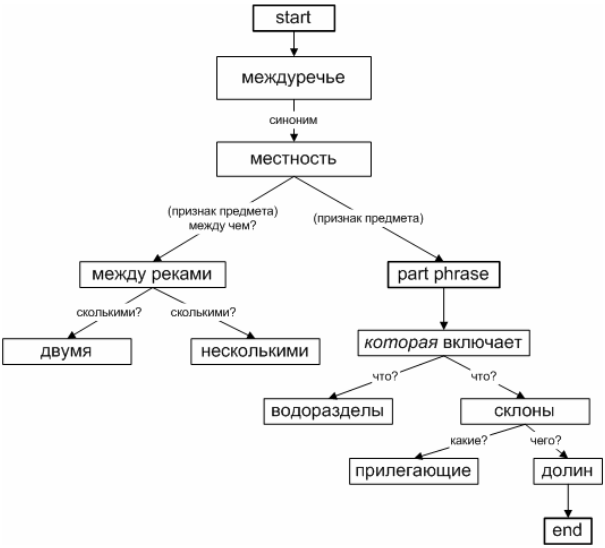

Рис. 1. Пример построения дерева, помеченного вопросами и ответами

4. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

Система состоит из набора функциональных модулей, представленных на рис. 2.

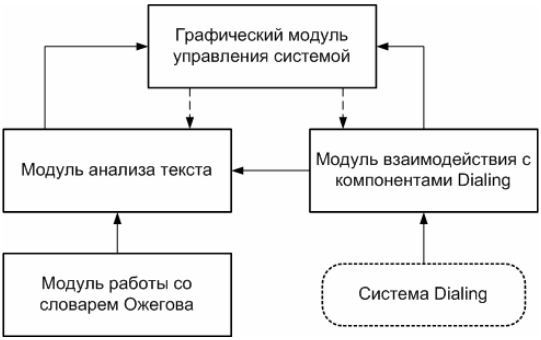

Рис. 2. Архитектура системы

Графический модуль управления системой выполняет функции общей координации работы системы. В этот модуль поступают все действия, произведённые пользователем, а также уведомления о внутренних событиях системы. В соответствии с полученной информацией этот модуль активизирует другие подсистемы для выполнения необходимой операции.

Модуль анализа текста является ядром системы и выполняет функции по генерации грамматических предикатов и построения деревьев с пометками по словарной статье из словаря Ожегова. Данный модуль взаимодействует с *модулем работы со словарем Ожегова* и с *модулем взаимодействия с компонентами Диалинг*.

Модуль взаимодействия с компонентами Диалинг. Задачей данного модуля является инициализация компонентов системы Диалинг и предоставление интерфейсов для основных функций, реализованных в этих модулях.

Модуль работы со словарем Ожегова обеспечивает доступ системы к словарю Ожегова, его основной задачей является получение словарной статьи из словаря.

Система реализована в среде Microsoft Visual Studio 6.0 на языке C++ с использованием MFC (Microsoft Foundation Classes). Такой выбор объясняется следующими преимуществами:

- язык C++ является одним из наиболее гибких языков в настоящее время;
- в связи с большой трудоёмкостью задачи, требуется наиболее производительное решение с большими возможностями к оптимизации.

4.1. Пользовательский интерфейс

Общий вид пользовательского интерфейса системы представлен на рис. 3.

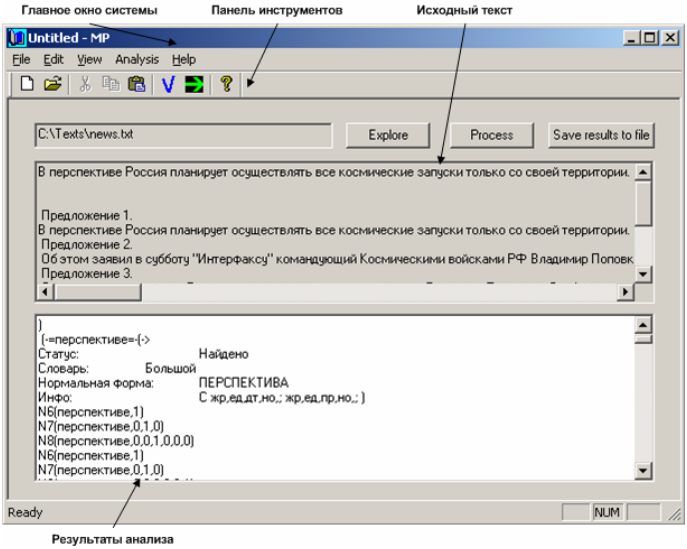

Рис. 3. Пользовательский интерфейс системы

На панели инструментов находятся кнопки, управляющие системой.

- Создание нового текста
- Загрузка текста из файла
- Вырезка выделенного участка текста с помещением в буфер обмена
- Копирование выделенного участка текста в буфер обмена
- Вставка текста из буфера обмена
- Установка параметров анализа
- Запуск анализа
- Информация о системе

Отметим, что все кнопки панели инструментов продублированы в меню. Кнопки входят в набор стандартных элементов управления большинства современных систем, работающих с текстом, поэтому

описание их работы приводить не будем; в системе они обладают стандартной функциональностью.

4.2. Анализ текстов

Перед запуском анализа текста пользователю необходимо выбрать файл, содержащий текст, который нужно проанализировать. Это можно сделать несколькими способами: нажав кнопку «Explore», через меню или через панель инструментов, нажав  . Текст загрузится в верхнее текстовое окно и для удобства будет разбит по предложениям (рис. 3).

Далее, для непосредственного запуска анализа текста пользователю необходимо нажать кнопку «Process» (то же самое действие производит  на панели инструментов, также анализ текста можно запустить из меню). Результаты анализа будут выведены в нижнем текстовом окне и будут содержать следующее.

- Информацию, полученную из системы Диалинг:
  - статус: найдено/не найдено слово в словаре системы;
  - словарь, в котором найдено слово;
  - лемма или нормальная форма слова;
  - набор граммем для данного слова.
- Грамматические предикаты, соответствующие данной части речи.
- Определение анализируемого слова из словаря Ожегова.
- Данные для построения дерева, помеченного вопросами и ответами.

Полученные результаты можно сохранить в текстовый файл посредством нажатия на кнопку «Save results to file» или через меню.

4.3. Установка параметров анализа

При нажатии на кнопку  появляется диалоговое окно, в котором устанавливаются параметры для анализа текстов (рис. 4). Пользователю предоставляется возможность выбрать, какие грамматические предикаты нужно генерировать по частям речи. Опция «Generate all» автоматически выбирает все предикаты. Также есть возможность включать/отключать вывод определения из словаря Ожегова и генерирование данных для построения дерева, помеченного вопросами и ответами.

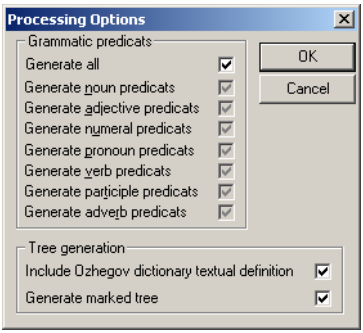

Рис. 4. Форма установки параметров анализа

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы были изучены и проанализированы различные системы автоматической обработки текста и алгоритмы, на которых эти системы базируются, их возможности, достоинства и недостатки. Были подробно изучены функциональные возможности и интерфейсы, предоставляемые компонентами системы Диалинг, при этом пришлось преодолеть трудности, связанные с неполной и, к тому же, не всегда достоверной документацией интерфейсов. В связи с этим, приходилось опробовать разные варианты работы с внешними компонентами, а также исследовать их с привлечением некоторых дополнительных инструментов, что существенно усложнило процесс разработки.

Также, ввиду отсутствия толковых словарей в формате, удобном для программного доступа, словарь Ожегова был переведен из плохо структурированных текстовых документов в базу, к которой легко обращаться программно.

Основным результатом работы является исследовательская программная система, использующая два различных подхода для анализа текстов на естественном языке:

- построение лексических функций и грамматических предикатов;
- представление предложений с помощью деревьев с пометками, построенных по словарной статье словаря Ожегова.

Результаты работы предполагается использовать в самых разных областях: от фундаментальной лингвистики до прикладного уровня, например, в

интеллектуальных поисковых системах, в системах автоматического резюмирования и т.д.

На данный момент еще не разработан алгоритм для сравнения и сопоставления помеченных деревьев, построенных по словарной статье из словаря Ожегова. В перспективе планируется реализация такого алгоритма, что позволит проверить на практике эффективность подхода.

СПИСОК ЛИТЕРАТУРЫ

1. **Мельчук И.А.** Опыт теории лингвистических моделей типа «Смысл⇔Текст». — М.: Наука, 1974.
2. **Сокирко А.В.** Реализация первичного семантического анализа в системе Диалинг // Тр. Междунар. семинара «Диалог'2000» по компьютерной лингвистике и ее приложениям, 1–5 июня 2000 г., Протвино.
3. **Панкратов Д.В., Гершензон Л.М.** Описание синтаксического анализа в системе Диалинг. Техн. документация по сист. Диалинг. — М., 1999.
4. **Леонтьева Н.Н. и др.** Семантический компонент в системах автоматического понимания текстов // Обзорная информация. — М.,1982. — Вып. 6.
5. **Леонтьева Н.Н.** Этапы информационного анализа естественного текста // Международный форум по информации и документации. — М., 1987. — Т. 12, № 4. — С. 8–14.
6. **Леонтьева Н.Н.** Система французско-русского автоматического перевода (ФРАП): лингвистические решения, состав, реализация // МП и ПЛ. Проблемы создания системы автом. перевода: Сб. научн. трудов МГПИИЯ им. М. Тореза. — М.,1987. — Вып. 271. — С. 6–25.
7. **Кудряшова И.М.** О семантическом словаре в системе ФРАП // МГПИИЯ им. М. Тореза: Сб. научн. трудов. — М., 1986. — Вып. 271.
8. **Леонтьева Н.Н.** ПОЛИТекст: информационный анализ политических текстов: Сб. НТИ. — 1995. — Сер. 2, № 4.
9. **Батура Т.В.** Представление смысла текста на естественном языке и его лексический анализ // Технологии Microsoft в информатике и программировании. — Новосибирск, 2004. — С. 88–90.
10. **Батура Т.В., Еркаева О.Н., Мурзин Ф.А.** К вопросу об анализе текстов на естественном языке // Новые информационные технологии в науке и образовании. — Новосибирск, 2003. — С. 7–58.
11. **Батура Т.В., Мурзин Ф.А.** Логические методы представления смысла текста на естественном языке // Новые информационные технологии в науке и образовании. — Новосибирск, 2003. — С. 59–111.
12. **Батура Т.В., Корда О.В.** Программные средства для анализа текстов на естественном языке // МНСК-XLII, 15–19 апреля 2004 г., Новосибирск.