Обеспечение содержательного доступа к информационным ресурсам по компьютерной лингвистике

Ю.А. Загорулько,
кандидат технических наук

Е.Г. Соколова,
кандидат филологических наук

И.С. Кононенко

Г.Б. Загорулько

О.И. Боровикова

В статье рассматривается интернет-портал знаний, обеспечивающий систематизацию и интеграцию знаний и информационных ресурсов по компьютерной лингвистике, а также содержательный доступ к ним (поиск информации и навигацию в терминах предметной области портала). Для целостного представления знаний и информационных ресурсов по компьютерной лингвистике их систематизация и структуризация выполнены на основе онтологии. Благодаря этому вся информация на портале представлена в виде сети взаимосвязанных информационных объектов.
Ключевые слова: портал знаний, компьютерная лингвистика, онтология, информационные ресурсы, содержательный доступ

Введение

В связи с постоянно растущими потребностями в средствах автоматической обработки документов и естественно-языковых, в том числе речевых, интерфейсах возникает необходимость в эффективном доступе не только к публикациям, описывающим методы и подходы, разработанные в лингвистике, но и к разного рода словарям, программным компонентам

и алгоритмам, реализующим различные задачи обработки текста и звучащей речи. В настоящее время в сети Интернет представлен большой объём знаний и информационных ресурсов по этой тематике, однако доступ к ним значительно затруднён, так как они систематизированы лишь частично и к тому же рассредоточены по различным интернет-сайтам, каталогам и электронным архивам.

Для устранения подобных проблем создаются специальные интернет-ресурсы, которые выполняют информационную поддержку разнообразных научных и тематических сообществ. Самым известным ресурсом такого рода, имеющим отношение к компьютерной лингвистике (КЛ), является англоязычный каталог LINGUIST List (http://linguistlist.org/), созданный для общения и обмена знаниями между лингвистами. Он содержит информацию о публикациях, персоналиях, научных учреждениях и других организациях лингвистического направления, грантах, конкурсах, проектах, фондах и источниках финансирования, а также о научных мероприятиях в лингвистической сфере деятельности. Кроме того, LINGUIST List предоставляет возможность поиска ресурсов по таким параметрам, как страна, язык, раздел лингвистики.

Из других зарубежных разработок стоит отметить созданный в Германском Исследовательском Центре Искусственного Интеллекта (DFKI) информационный портал «Language Technology World» (http://www.lt-world.org/). Тематические разделы этого портала содержат информацию о лингвистических технологиях, продуктах и информационных системах в области обработки естественного языка, а также о проектах, организациях, персонах. В основу портала положена онтология, благодаря чему возможно установление связей между его разделами. К сожалению, на этом портале практически отсутствует информация об исследованиях, проводимых в России.

К российским аналогам LINGUIST List можно отнести научно-образовательный портал «Лингвистика в России: ресурсы для исследователей» (http://uisrussia.msu.ru/linguist/index.jsp) и сайт «Российская лингвистика (RUSLING)» (http://rusling.narod.ru), который разрабатывается в Отделении лингвистических исследований ВИНИТИ РАН.

Портал «Лингвистика в России» содержит иерархически организованный каталог ссылок на наиболее значимые лингвистические ресурсы и позволяет осуществлять навигацию по разделам портала с помощью иерархических связей внутри разделов и по ссылкам на связанные с ними области (разделы). Тематические категории данного портала представлены разделами по компьютерной, теоретической и прикладной лингвистике и их приложениям (смежным областям), а также разделами, посвящёнными русскому языку, языкам мира и народов РФ.

Портал «Российская лингвистика» предлагает лингвистам «информационную карту» для поиска информации об организациях, научных исследованиях и публикациях, лингвистических ресурсах и персоналиях. Он содержит обширный каталог ссылок на словари и корпуса текстов для различных языков (в том числе славянских), а также сведения о российских лингвистах, предоставляя возможность их поиска не только по алфавиту, но и по области и объекту (языку) исследования.

Примером специализированного тематического ресурса по КЛ является российский сайт «Речевые технологии» (http://speech-soft.ru/), на котором представлена информация, охватывающая прикладные аспекты данного направления (технологии, программные средства, коллективы разработчиков, конкретные системы и т.п.).

Как правило, научно-практические проекты, разрабатываемые в рамках описанных выше подходов, направлены либо на описание и сохранение общей лингвистической информации, либо на представление информации о каком-то одном разделе лингвистики, но не для интеграции ресурсов по компьютерной лингвистике и обеспечения доступа к ним широкому кругу пользователей.

Для решения этих проблем в рамках предложенного нами подхода к построению специализированных интернет-порталов знаний [1] разработан портал знаний по компьютерной лингвистике. Как информационный ресурс указанный портал обеспечивает следующие возможности:

панорамную характеристику научного направления «компьютерная лингвистика» через представление используемых в нём терминов и понятий, объектов и методов исследования, научных результатов, а также участников научной деятельности в рамках этого направления (персон, групп, сообществ и других организаций, вовлечённых в процесс исследования);

интеграцию доступных информационных ресурсов по компьютерной лингвистике в единое информационное пространство;

содержательный доступ к систематизированным знаниям и данным, относящимся к компьютерной лингвистике, т.е. возможность поиска и получения информации в терминах предметной области портала, а также удобную навигацию по всему информационному пространству портала, базирующуюся на модели его предметной области;

информационную поддержку пользователей, т.е. анонсирование разного рода событий и мероприятий, касающихся данного научного направления.

1. Информационная модель портала знаний по КЛ

В качестве концептуальной основы информационной модели портала знаний выбрана онтология [2], с помощью которой можно достаточно просто обеспечить унифицированное представление и хранение знаний и информационных ресурсов по компьютерной лингвистике, а также содержательный доступ к ним.

Онтология портала обеспечивает представление понятий, необходимых для описания как научной деятельности и научного знания в целом, так и конкретной научной дисциплины в частности. Поэтому онтология портала включает универсальные онтологии научной деятельности и научного знания [3], а также онтологию предметной области.

Универсальные онтологии не зависят от предметной области (ПО) и могут использоваться практически в любом портале научных знаний, независимо от его конкретной тематики. В связи с этим указанные онтологии выделены в качестве базовых (рис. 1). Рассмотрим их подробнее.

Онтология научной деятельности является онтологией верхнего уровня и включает базовые понятия, относящиеся к **организации** научно-

исследовательской деятельности, такие, как *Персона*, *Организация*, *Событие*, *Деятельность*, *Публикация*, которые используются для описания участников научной деятельности, мероприятий, научных программ и проектов, различного типа публикаций, а также *Географическое место*. В эту онтологию также включено понятие *Информационный ресурс*, которое служит для описания информационных ресурсов, представленных в сети Интернет.

Онтология научного знания, по своей сути, является метаонтологией. Она содержит метапонятия и отношения, задающие структуры для описания рассматриваемой предметной области, такие, как *Раздел науки, Предмет исследования, Объект исследования, Метод исследования, Научный результат*, позволяющие выделить в данной науке значимые разделы и подразделы, задать типизацию предметов, объектов и методов исследования, описать результаты научной деятельности.

Понятия базовых онтологий связаны между собой ассоциативными отношениями (см. рис. 1), выбор которых осуществлялся не только исходя из полноты представления моделируемой области знаний портала, но и с учётом удобства навигации по его информационному пространству и поиска информации.

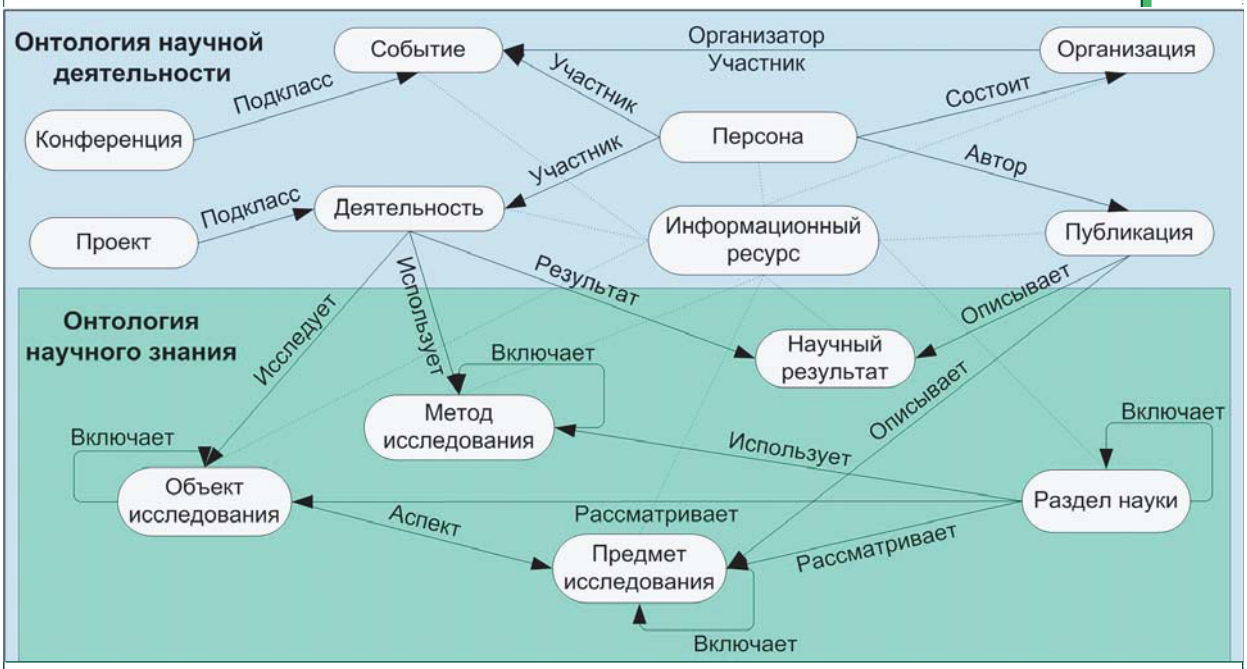


Рис. 1. Базовые онтологии портала

В качестве онтологии предметной области портал включает онтологию компьютерной лингвистики, фрагмент которой представлен на рис. 2. Понятия этой онтологии являются реализациями метапонятий онтологии научного знания и организованы в пять иерархий «общее-частное», каждая из которых соответствует одному из перечисленных выше метапонятий. Все эти иерархии связаны между собой посредством ассоциативных отношений, часть которых наследуется из базовых онтологий, а часть отражает специфику данной предметной области.

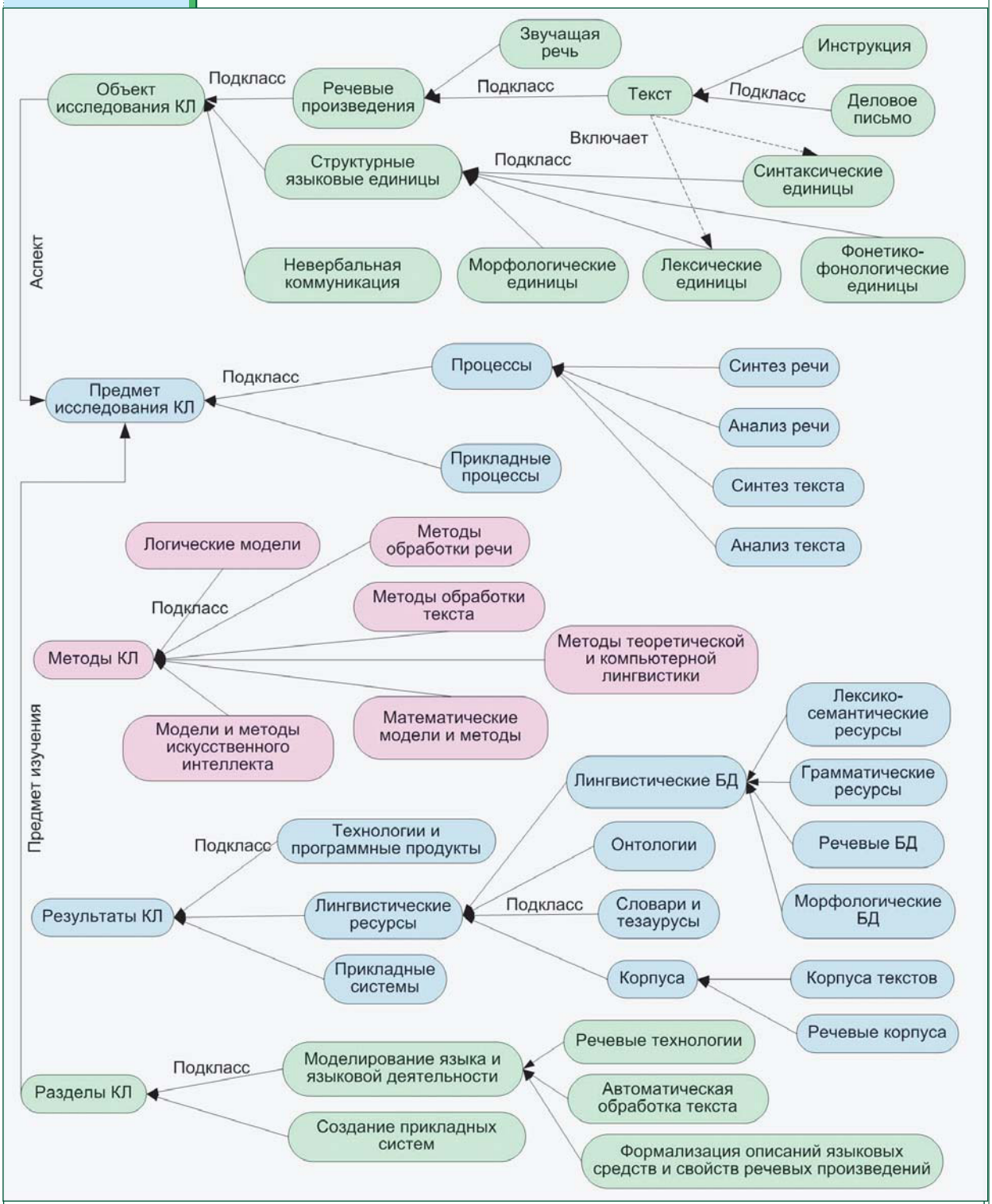


Рис. 2. Фрагмент онтологии компьютерной лингвистики

Рассмотрим онтологию компьютерной лингвистики подробнее.

В качестве **базовых объектов** исследования КЛ предложено рассматривать *Речевое произведение* (РП) как объективную форму существования и использования естественного языка, *Структурные языковые единицы* в составе РП, соответствующие различным языковым уровням: предложения, словосочетания, слова, морфемы, звуки и интонационные единства, а также *Невербальную коммуникацию*.

Класс понятий РП в зависимости от формы (графической или звуковой) представлен в иерархии двумя подклассами: *Текст* и *Звучащая речь*. Выделяемые в РП *Языковые единицы* сгруппированы в соответствии с языковыми уровнями в классы: *Синтаксические единицы, Лексические единицы, Морфологические единицы* и *Фонетико-фонологические единицы*. Для представления связи между целостными РП и их структурными единицами используется отношение «Включение».

Предметом исследования в КЛ являются *Процессы и задачи*, связанные с функционированием языковых единиц в коммуникации, и *Прикладные процессы и задачи*, имеющие практическую ценность, отвечающие определённому социальному запросу. Иерархия предметов исследования связана ассоциативным отношением «Аспект» с иерархией объектов исследования и отношением «Предмет исследования раздела науки» с иерархий разделов науки.

Иерархия методов исследования служит для систематизированного описания инструментов исследования, применяемых в компьютерной лингвистике. В этой иерархии были выделены подклассы понятий *Методы теоретической и компьютерной лингвистики, Методы обработки текста, Методы обработки речи, Модели и методы искусственного интеллекта, Логические модели* и др.

В основе Иерархии разделов КЛ лежит классификация базовых теоретических и прикладных направлений компьютерной лингвистики. В качестве **главных** выделены разделы *Моделирование языка и языковой деятельности* (с подразделами *Автоматическая обработка текста (АОТ), Речевые технологии (РТ), Формализация описаний языковых средств и свойств речевых произведений*) и *Создание прикладных систем*. В зависимости от направления моделирования (анализ или синтез) в классе понятий *Автоматическая обработка текста* выделены соответствующие подклассы: *Понимание текста* и *Генерация текста*, а в классе *Речевые технологии — Распознавание речи* и *Синтез речи*. В зависимости от объекта обработки (текст или звучащая речь), *Прикладные системы* разделены на классы *Создание прикладных систем АОТ* и *Создание прикладных систем РТ*.

Иерархия Научных результатов служит для типизации и описания результатов научной деятельности. В этой иерархии выделены следующие классы: *Технологии и программные продукты, Прикладные системы, Лингвистические ресурсы*. Последний класс делится на такие классы: *Корпуса*, *Лингвистические БД, Онтологии, Словари и тезаурусы*.

Таким образом, вводя формальные описания проблемной и предметной области в виде понятий и отношений между ними, онтология портала задаёт структуры для представления реальных объектов и связей между ними.

В соответствии с принятой нами моделью данные на портале представлены в виде множества разнотипных информационных объектов и связей между ними. *Информационный объект* (ИО) — это структурированная совокупность данных, представляющая собой

описание некоторого объекта из выбранной области знаний или релевантного для неё информационного ресурса. Каждый ИО соответствует некоторому понятию онтологии и имеет заданную им структуру. Между конкретными информационными объектами могут существовать связи, семантика которых определяется отношениями, заданными между соответствующими понятиями онтологии.

2. Информационное содержание портала знаний по КЛ

Информационное содержание (контент) портала включает как знания общего характера, так и конкретные знания о реальных объектах и информационных ресурсах, систематизированные в соответствии с онтологиями портала.

В контенте портала КЛ представлены, прежде всего, знания об основных разделах компьютерной лингвистики, о её предметах и объектах исследования, используемых в ней моделях и методах. Кроме этого, пользователи портала могут найти информацию о выполняемой в области компьютерной лингвистики научной деятельности. В первую очередь, это информация об учёных, исследовательских группах и организациях и их деятельности. Так, например, при просмотре информации о «Группе речевых исследований при кафедре теоретической и прикладной лингвистики филологического факультета МГУ» можно увидеть список исследователей, занятых в деятельности этой группы, а также определить её место в структурной иерархии этого подразделения в рамках университета. Направление работ группы представлено такими разделами КЛ, как *Речевые технологии*, *Создание прикладных систем РТ* и *Формализация описаний языковых средств и свойств речевых произведений*. Кроме того, из описания группы можно перейти на проект ISABASE, в котором она участвовала, и на её сайт, являющийся основным информационным ресурсом этой группы.

В деятельности организаций и исследователей особое место занимают научные и коммерческие проекты, в рамках которых создаются лингвистические знания и ресурсы. Результаты этой деятельности находят отражение в публикациях — монографиях, статьях, материалах конференций и семинаров, отчётах и других текстовых ресурсах, доступ к которым предоставляется порталом. Например, на портале можно найти информацию о монографии Д. Журавского и Дж. Мартина «Speech and Language Processing. An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition», учебнике Р. Миткова «The Oxford handbook of computational linguistics», монографии Н.Н. Леонтьевой «Автоматическое понимание текста: системы, модели, ресурсы» и других публикациях.

Портал обеспечивает доступ и к информационным ресурсам, представляющим непосредственные результаты деятельности организаций и отдельных исследователей, полученные в ходе выполнения научных и коммерческих проектов, а именно: технологии, программные продукты, прикладные системы, лингвистические ресурсы: словари, корпуса (текстов и речи) и лингвистические БД. Для организации более эффективного доступа

к таким ресурсам в контенте представлена информация о различных аспектах их разработки: организациях, персонах и проектах, с которыми связано их появление, а также о таких содержательных характеристиках ресурсов, как отнесённость к разделу науки, объекту или предмету исследования, методам исследования. Эта информация связывает ресурсы с остальными данными и знаниями, представленными в контенте портала, что позволяет пользователю выделить группы ресурсов, созданные, например, в ходе осуществления некоторой исследовательской деятельности (гранта, проекта, конкурса) или с использованием определённого класса методов исследования. Например, при просмотре информации о лингвистическом ресурсе «Речевой корпус RuSpeech» пользователь может заметить, что тематика научного результата объединяет разделы КЛ *Речевые технологии* и *Создание корпусов*. А при дальнейшем переходе к просмотру описания проекта RuSpeech, в рамках которого создавался речевой ресурс, можно увидеть информацию о других результатах и публикациях этого проекта.

Важным компонентом информационного контента портала является описание интернет-ресурсов, к которым относятся сайты организаций, конференций, проектов, порталы и каталоги, а также отдельные страницы с материалами графического, мультимедийного или текстового типа. Как было сказано выше, каждый интернет-ресурс, представленный на портале, соответствует такому понятию онтологии, как Информационный ресурс. Описание отдельного ресурса включает экземпляр данного понятия и набор экземпляров отношений, связывающих его с другими объектами, представляющими организации, персоны, публикации, события и т.д.

3. Обеспечение доступа к ресурсам по компьютерной лингвистике

Основное назначение рассматриваемого портала знаний — обеспечить содержательный доступ к систематизированным знаниям и информационным ресурсам по компьютерной лингвистике. Доступ к знаниям и данным портала осуществляется путём навигации по дереву понятий онтологии и контенту портала (см. рис. 3), а также через развитые средства содержательного поиска.

3.1. Навигация по контенту портала

Для конечного пользователя данные на портале представлены в виде множества связанных информационных объектов. При навигации по порталу обеспечивается возможность выбора ИО, относящихся к интересующему пользователя понятию, просмотра и фильтрации списков выбранных ИО, навигации по конкретным ИО, а также просмотра описания выбранного информационного ресурса.

Список ИО отображается в виде страницы, содержащей набор ссылок на эти объекты. Для больших списков формируется составная страница, включающая список страниц с элементами навигации по этому списку.

Вся информация о конкретном объекте и его связях отображается в виде HTML-страницы (рис. 4), формат и наполнение которой зависят от свойств понятия, экземпляром которого является данный объект, и заданного для него шаблона визуализации. При этом объекты, связанные с данным объектом, представляются на его странице в виде гиперссылок, по которым можно перейти к их детальному описанию.

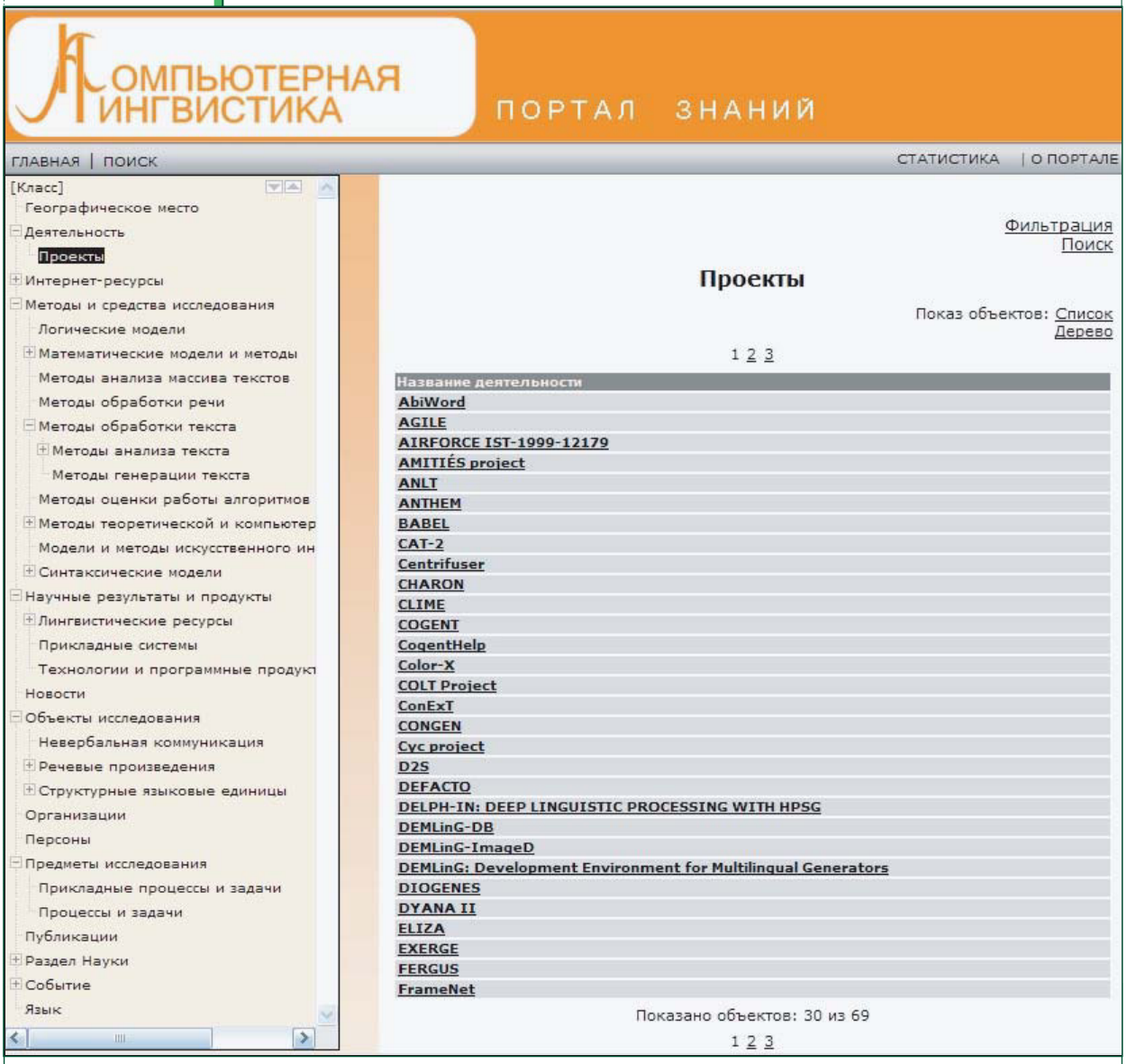
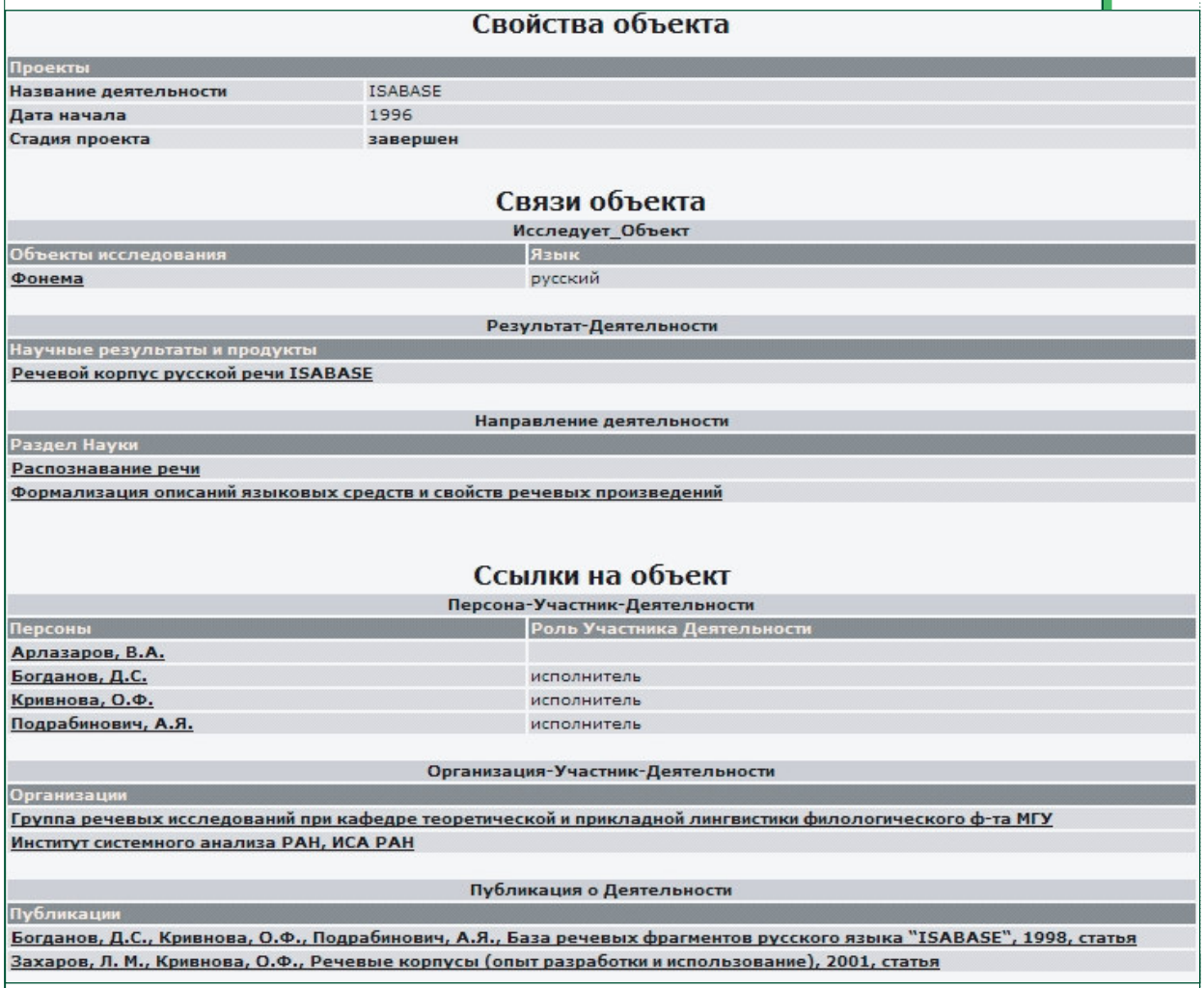


Рис. 3. Навигация по порталу знаний

Таким образом, навигация по данным портала представляет собой процесс перехода от одних информационных объектов к другим по заданным между ними связям.

Например, при просмотре информации о конкретном проекте (см. рис. 4) мы можем видеть значения его атрибутов и его связи с другими объектами. Используя представленные связи в качестве элементов навигации, можно перейти к просмотру подробной информации о научных результатах, полученных в ходе выполнения проекта, об участниках проекта, публикациях о нём и т.п.



Свойства объекта

Проекты	
Название деятельности	ISABASE
Дата начала	1996
Стадия проекта	завершен

Связи объекта

Исследует_Объект

Объекты исследования	Язык
Фонема	русский

Результат-Деятельности

Научные результаты и продукты
Речевой корпус русской речи ISABASE

Направление деятельности

Раздел Науки
Распознавание речи
Формализация описаний языковых средств и свойств речевых произведений

Ссылки на объект

Персона-Участник-Деятельности

Персоны	Роль Участника Деятельности
Арлазаров, В.А.	
Богданов, Д.С.	исполнитель
Кривнова, О.Ф.	исполнитель
Подрабинович, А.Я.	исполнитель

Организация-Участник-Деятельности

Организации
Группа речевых исследований при кафедре теоретической и прикладной лингвистики филологического ф-та МГУ
Институт системного анализа РАН, ИСА РАН

Публикация о Деятельности

Публикации
Богданов, Д.С., Кривнова, О.Ф., Подрабинович, А.Я., База речевых фрагментов русского языка "ISABASE", 1998, статья
Захаров, Л. М., Кривнова, О.Ф., Речевые корпусы (опыт разработки и использование), 2001, статья

Рис. 4. Представление информационного объекта и его связей

При переходе по конкретной связи любого информационного объекта мы можем получить достаточно большой список объектов (например, список людей, работающих в некоторой организации). В связи с этим был введён механизм фильтрации списков информационных объектов, который позволяет, например, отфильтровать множество публикаций как по дате публикации, так и по описываемому научному результату или объекту исследования.

3.2. Поиск в терминах предметной области портала

При поиске информации пользователю предоставляется возможность задания запроса в терминах предметной области портала. При этом пользователь должен выбрать понятие, к которому относятся искомые информационные объекты, и определить ограничения, которым должны удовлетворять атрибуты выбранного понятия и его связи с другими понятиями.

Ограничения на отдельные атрибуты интерпретируются как конъюнкция условий. Допустимые ограничения для атрибута зависят от типа его значений. Так, например, для атрибутов

типа «integer» и «date» задаётся точное значение или допустимый интервал значений.

Пользователю также предоставляется возможность задать условия на значения атрибутов объектов, связанных с искомым объектом. При этом могут быть заданы ограничения и на значения атрибутов соответствующих отношений.

Например, для получения ответа на вопрос «Найти проекты, выполнявшиеся после 1995 года, в которых принимал участие В.Л. Арлазаров и исследовались русские фонемы» пользователь должен выбрать в дереве онтологии понятие «Проект», а затем в автоматически сгенерированной поисковой форме задать ограничения на значения соответствующих атрибутов объектов и отношений. В результате будет сформирован следующий запрос:
Понятие *«Проект»*
Атрибут *«Дата начала»(>=1995)*
Отношение *«Исследует_Объект»*
Атрибут отношения *«Язык» = «русский»*
Понятие *«Объект исследования»*
Атрибут *«Название» = «Фонема»*
Отношение *«Персона-Участник-Деятельности»:*
Понятие *«Персона»*
Атрибут *«Фамилия» = «Арлазаров»*
Атрибут *«Инициалы» = «В.Л.»*

Заключение

В статье описан специализированный интернет-портал, обеспечивающий содержательный доступ к знаниям и информационным ресурсам по компьютерной лингвистике.

Портал представляет знания об основных разделах компьютерной лингвистики, о её предмете и объектах исследования, используемых в ней моделях и методах, разработанных системах, алгоритмах и лингвистических ресурсах, содержит информацию об учёных, сообществах, организациях, вовлечённых в процесс исследований по компьютерной лингвистике, о выполняемых проектах в этой области. Пользователи портала имеют доступ не только к текстовым ресурсам по КЛ, но и к ресурсам, представляющим реальные прикладные системы, технологии и программные продукты для обработки естественного языка, словари и лингвистические базы данных.

Благодаря тому, что систематизация и структуризация знаний и данных по компьютерной лингвистике выполнена на основе онтологии, вся информация на портале представлена в виде сети взаимосвязанных информационных объектов. Доступ к знаниям и данным портала осуществляется путём навигации по дереву понятий онтологии и его информационному пространству, а также через средства содержательного поиска.

Портал знаний по компьютерной лингвистике функционирует и доступен по адресу http://uniserv.iis.nsk.su/cl/. В планах авторов — дальнейшее развитие онтологии компьютерной лингвистики, сбор и интеграция в контент портала новых лингвистических знаний и информационных ресурсов.

Литература

1. Загорулько Ю.А., Боровикова О.И., Загорулько Г.Б. Организация содержательного доступа к информационным ресурсам на основе онтологий // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Тр. 9-й Всероссийской научной конф. RCDL'2007. Переславль-Залесский: Изд-во «Университет города Переславля», 2007. Т.1. С. 217–224.

2. Guariano N., Giaretta P. Ontologies and Knowledge Bases. Towards a Terminological Clarification // Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing. Amsterdam: IOS Press, 1995. P. 25–32.

3. Загорулько Ю.А., Боровикова О.И. Технология построения онтологий для порталов знаний по гуманитарным наукам // Труды Всероссийской конференции с международным участием «Знания-Онтологии-Теории» (ЗОНТ-07). Новосибирск, 2007. Т.1. С. 191–200.

Загорулько Юрий Алексеевич,
кандидат технических наук, заведующий лабораторией
Института систем информатики им. А.П. Ершова СО РАН.
e-mail: zagor@iis.nsk.su.

Е.Г. Соколова,
кандидат филологических наук,
доцент Российского государственного гуманитарного университета, Москва.

И.С. Кононенко,
научный сотрудник ИСИ СО РАН.

Г.Б. Загорулько,
научный сотрудник ИСИ СО РАН.

О.И. Боровикова,
младший научный сотрудник ИСИ СО РАН.