На правах рукописи

Архангельский Тимофей Александрович

Принципы построения морфологического парсера для разноструктурных языков

Специальность 10.02.21 — прикладная и математическая лингвистика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата филологических наук

Москва – 2012

Работа выполнена на отделении теоретической и прикладной лингвистики филологического факультета Федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Московский государственный университет имени М. В. Ломоносова»

НАУЧНЫЙ РУКОВОДИТЕЛЬ:

член-корреспондент РАН,
доктор филологических наук,
профессор
Плунгян Владимир Александрович

ОФИЦИАЛЬНЫЕ ОППОНЕНТЫ:

Рябцева Надежда Конатантиновна
доктор филологических наук,
ведущий научный сотрудник
(Институт языкознания РАН,
зав. сектором прикладного языкознания)

Иомдин Леонид Лейбович
кандидат филологических наук,
старший научный сотрудник
(Институт проблем передачи информации,
ведущий научный сотрудник
лаборатории компьютерной лингвистики)

ВЕДУЩАЯ ОРГАНИЗАЦИЯ:

ФГБУН «Институт проблем информатики Российской академии наук»

Защита диссертации состоится 26 декабря 2012 г. в ____ часов
на заседании диссертационного совета Д 501.001.24 при МГУ им. М. В. Ломоносова
по адресу: 119991, ГСП-1, Москва, Ленинские горы, МГУ,
1-й корпус гуманитарных факультетов, филологический факультет

С диссертацией можно ознакомиться в Научной библиотеке МГУ им. М. В. Ломоносова.
Автореферат разослан _____ 2012 г.

Ученый секретарь
диссертационного совета (А. М. Белов)

Общая характеристика работы

Данное исследование посвящено изучению и решению проблем, которые возникают при морфологической разметке языковых корпусов. В работе предлагается способ формализованного описания грамматики и лексики языков, охватывающий широкий круг морфологических явлений и позволяющий использовать его при создании корпусов разноструктурных языков. Формат и построенный на его основе морфологический парсер были успешно использованы при создании ряда корпусов.

Объектом исследования являются проблемы и специфические задачи, возникающие при создании крупных корпусов языков, обладающих сложной морфологической системой.

В настоящий момент благодаря развитию компьютерных технологий электронные корпуса языков стали повсеместно использоваться как инструмент лингвистического исследования, а корпусная лингвистика за последние два десятилетия стала одной из важных областей не только прикладной, но и теоретической лингвистики. Одной из самых важных задач при составлении корпуса языка является создание так называемого морфологического анализатора, или парсера — компьютерной системы автоматического морфологического анализа языка. С помощью морфологического парсера всем словоформам из текстов на каком-либо языке, образующих корпус, ставится в соответствие начальная (словарная) форма, набор грамматических характеристик и, возможно, другая информация, по которой пользователи корпуса смогут осуществлять поиск. Именно наличие такого рода разметки делает корпус ценным инструментом лингвистического исследования. Если корпус текстов относительно невелик (десятки или сотни тысяч словоупотреблений), такую разметку можно внести в текст вручную, без помощи специальных средств. Однако выполнить разметку большого корпуса без парсера практически невозможно — этим объясняется его исключительная

важность при создании корпусов.

Если создание парсера для морфологически бедного языка (по крайней мере, для языка с бедным словоизменением) — например, английского или французского — не представляет больших проблем, создание парсера для языков с богатой морфологией и множеством нетривиальных морфологических явлений может быть сопряжено со значительно большими трудностями. Создание такого парсера является сложной задачей, требующей больших затрат времени и ресурсов.

Для некоторых существующих корпусов с морфологической разметкой были специально написаны парсеры, способные анализировать тексты на одном языке. Однако, учитывая постоянно увеличивающееся количество создаваемых корпусов, массовое применение такого решения представляется довольно неэффективным. Каждый язык требует долгой совместной работы программистов и лингвистов для создания с нуля очередной системы морфологического анализа. Между тем, с технической точки зрения эти системы имеют много общего — и этот факт можно было бы выгодно использовать, выделив эту общую часть в отдельный продукт, который можно многократно применять при создании парсеров конкретных языков. Более того, эта общая часть может иметь вид цельной программной системы — *универсального парсера*, шаблонной программы, требующей для работы в качестве парсера некоторого языка только специальным образом составленное формальное описание этого языка. При этом не только достигается экономия времени составителей корпусов, но и отпадает необходимость в использовании труда программиста.

В настоящем исследовании предлагается способ именно такого формализованного описания языка, учитывающий множество морфологических явлений типологически разных языков и специфических проблем, возникающих при создании крупных языковых корпусов.

Актуальность исследования. Несмотря на успехи, достигнутые в области корпусной лингвистики за последние десятилетия, для подавляющего большинства языков всё ещё не созданы корпуса (и далеко не для всех языков эта задача является тривиальной), поэтому данная область имеет огромный потенциал развития. В настоящий момент существует небольшое количество универсальных парсеров (или, по крайней мере, парсеров, претендующих на возможность их использования для анализа множества типологически разных языков). Однако все они обладают недостатками, по той или иной причине затрудняющими их использование для разметки крупных корпусов. В то же время создание пригодного для практического применения универсального парсера в настоящий момент является крайне важной задачей, решение которой позволит упростить и ускорить создание таких корпусов.

Цели исследования продиктованы описанными выше запросами современной корпусной лингвистики. Ниже они перечислены в порядке убывания важности:

1. Выявить круг проблем и задач, возникающих при создании крупных корпусов с морфологической разметкой.

2. Учитывая полученные результаты, разработать формат описания лексики и грамматики языка, пригодный для использования универсальным парсером. Требования к этому формату таковы:

- Формат должен обладать достаточными средствами для описания широкого спектра морфологических явлений, представленных в типологически различных языках.
- Формат и использующий его универсальный парсер должны быть ориентированы на разметку большого количества текстов для корпусов (в частности, должна обеспечиваться достаточно высокая скорость морфологического анализа).
- Формат должен быть ориентирован, в частности, на работу с

письменными текстами, созданными на языках с кодифицированной орфографией; в нём должно учитываться исключительно графическое представление словоформ без обращения к их фонемному составу или какой бы то ни было глубинной структуре.

- Формат должен давать возможность описать словоизменение, не вынуждая при этом пользователя указывать разбиение словоформ на морфемы или составлять отдельные словари морфем.
- Формат должен быть в том числе доступен для использования лингвистами без навыков программирования; в случаях, когда без этого невозможно обойтись, необходимо прибегать к применению уже существующих и широко используемых средств вместо изобретения собственных аналогов (в первую очередь это касается языка регулярных выражений).
- Файлы в данном формате должен иметь достаточно простую структуру, чтобы, во-первых, лингвист мог самостоятельно вносить в них информацию с применением самых простых текстовых редакторов, доступных для любой операционной системы, а во-вторых, эти файлы могли легко обрабатываться с применением одного из языков программирования.

Предполагается, что при составлении корпуса нужно руководствоваться рядом правил, таких как теоретическая нейтральность или приоритет широты поисковых возможностей. Требования к формату являются следствием этих принципов и принципа экономии усилий при составлении описания языка.

3. Создать пилотную версию универсального парсера, способную использовать большинство элементов данного формата. Требования, предъявляемые к парсеру, таковы.

- Парсер должен быть способен, во-первых, достаточно быстро анализировать тексты на языке, описываемом передаваемыми ему файлами, а во-вторых, порождать парадигму каждой лексемы

описываемого языка для проверки правильности составления грамматического описания.

- При наличии необходимой информации в предоставляемом ему описании языка парсер должен производить не только грамматический анализ, но и глоссирование текста. Глоссирование должно осуществляться в соответствии с широко известными лейпцигскими правилами глоссирования, а в случаях, не регулируемых этими правилами, пользователю должна быть предоставлена возможность выбора способа глоссирования с учётом существующей практики.
- Парсер должен получать на вход тексты в обычном текстовом формате и выдавать размеченный текст в широко используемом формате XML, что позволит производить его дальнейшую обработку в других системах или непосредственное помещение их в корпус.

При этом парсер не должен выполнять никаких других операций, обрастая не свойственными ему функциями. В частности, в его задачи не входит снятие омонимии: каждой словоформе должны быть приписаны все разборы, возможные с точки зрения переданного парсеру описания. При необходимости такой парсер в качестве самодостаточного модуля может быть встроен в более крупную систему (например, включающую графический интерфейс для создания словаря и описания грамматики, интерфейс для ручной правки глоссированного текста или средства для снятия омонимии).

4. Применить созданные формат и парсер на практике, использовав их для разметки текстов на нескольких разноструктурных языках с дальнейшим использованием размеченных текстов в реальных корпусах.

На защиту выносятся следующие положения:

1. В условиях необходимости создания большого количества языковых корпусов с морфологической разметкой наиболее рациональной стратегией

является разработка таких инструментов, позволяющих работать с разноструктурными языками, как единый формат записи данных о грамматике и универсальный парсер.

2. Создание формата и системы со свойствами, пригодными для такого использования, теоретически и практически возможно.

Научная новизна исследования состоит в том, что впервые был создан формат описания грамматики со свойствами, перечисленными выше, и с помощью данного формата и парсера были впервые созданы корпуса нескольких разноструктурных языков.

Теоретическая значимость исследования состоит в том, что были изучены проблемы, возникающие при создании крупных корпусов с морфологической разметкой (в частности, при формализованном описании их грамматики), и были предложены варианты их решения.

Практическая значимость исследования состоит в том, что был разработан формат описания лексики и грамматики языка, обладающий рядом свойств, которые позволяют эффективно использовать его при создании корпусов. С помощью парсера, работающего с данным форматом, были созданы корпуса албанского, калмыцкого, лезгинского и осетинского языков, а в настоящий момент создаются корпуса новогреческого и бурятского языков. Результаты исследования могут использоваться для создания корпусов других языков.

Основным материалом исследования послужили данные (тексты и грамматические описания), обработанные в ходе создания корпусов албанского, калмыцкого, лезгинского и осетинского языков в 2011 г.

Апробация работы. Основные положения работы были представлены и

обсуждены на заседании кафедры теоретической и прикладной лингвистики филологического факультета МГУ, на рабочих семинарах отделения теоретической и прикладной лингвистики и в рабочих группах, занимающихся созданием перечисленных выше корпусов, а также опубликованы (см. список публикаций).

Структура работы. Работа состоит из введения, трёх глав, заключения, библиографии и приложения. Объём работы (без приложений и библиографии) составляет 182 страницы.

Краткое содержание работы

Во **введении (глава 1)** описываются предмет, цели и задачи исследования и их обоснование, а также даётся обзор существующих средств для автоматической морфологической разметки текстов.

Глава 2 является основной частью работы. В этой главе приводится полное описание предлагаемого формата UniParser с обоснованием выбора реализуемых в нём решений и соответствия их требованиям, предъявленным в главе 1.

В формате UniParser описание языка построено в целом по словесно-парадигматической модели: базовыми единицами описания являются лексема и парадигма, а не отдельные морфемы; впрочем, для описания языков с богатой агглютинативной аффиксацией предусмотрены специальные механизмы. Понятием морфемы в этом формате можно вообще не пользоваться — делить сложные показатели на морфемы и проводить лингвистически корректную границу между основой и словоизменительными показателями необходимо только в том случае, если пользователю нужна возможность глоссирования текстов.

Все сведения о языке содержатся в нескольких словарных файлах,

имеющих обычный текстовый формат (plain text) и кодировку UTF-8. В отдельных файлах хранится информация о лексике языка, о грамматике, о продуктивном словоизменении и т. д. Ниже приводится общий список файлов, за которым следует описание каждого из них.

Основные файлы:

- stems.txt — список лексем, для которых указываются основы, словоизменительные классы и любая другая лексическая информация, т. е. та информация, которая должна быть приписана каждой словоформе данной лексемы;
- paradigms.txt — список словоизменительных показателей, объединённых в парадигмы разных словоизменительных типов;
- derivations.txt — список продуктивных словообразовательных моделей;
- clitics.txt — список клитик, которые пишутся слитно с опорным словом или отделяются знаком, отличным от пробела;
- stem_conversion.txt — список правил, позволяющих автоматически получать одни алломорфы основ из других;
- periphrastic_forms.txt — список аналитических форм.

Вспомогательные файлы:

- punc.txt — список знаков препинания;
- ignore.txt — список символов, которые парсер должен игнорировать;
- equiv.txt — список пар символов, которые должны считаться эквивалентными при анализе текста.

Вспомогательные файлы имеют тривальную структуру: это просто списки элементов (например, знаков пунктуации), в которых на каждой строке находится один элемент. Основные файлы содержат более сложную информацию и поэтому должны иметь формат, позволяющий хранить структурированные данные. Эти файлы устроены по-разному, поскольку они предназначены для хранения информации разного типа, однако все они

основаны на известном формате YAML. Хотя наиболее часто применяющимся для хранения структурированных данных форматом является XML, YAML был выбран из-за его большей удобочитаемости и меньшей избыточности по сравнению с XML.

Из всех файлов обязательными для работы являются только stems.txt и paradigms.txt. Отсутствие остальных файлов означает либо то, что соответствующие механизмы не будут использоваться (как в случае со словообразованием), либо то, что будут использованы значения по умолчанию (как в случае со знаками препинания).

При описании языка используется только графическое представление слов, т. е. то, как слово выглядит в реальном тексте; в формате UniParser не используются фонемные или какие-либо «глубинные» представления слов, не совпадающие с графическим. Причина такого решения — обеспечение простоты теоретической базы, относительной независимости от тех или иных теоретических рамок и практическая направленность: формат предназначен в первую очередь для разметки текстов для крупного корпуса, написанных в том числе на языках с давно сложившейся орфографической нормой, а не составление полностью лингвистически корректного описания грамматики языка в формализованном виде. В этом мы следуем за грамматическим словарём Зализняка, где все основы и окончания приводятся только в графическом виде.

Основы и словоизменительные типы

В модели описания языка, принятой в формате UniParser, считается, что у каждой лексемы языка есть парадигма — множество словоформ, объединённых принадлежностью к одной лексеме, но, возможно, различающихся выраженными в них грамматическими значениями. У каждой лексемы имеется канонический представитель, называемый начальной формой, или леммой. Также мы считаем, что каждую словоформу языка — или, по крайней мере, подавляющее их большинство — можно условно разделить на основу и

флексию. В первом приближении под основой понимается та часть словоформы, которая не изменится при переходе к какой-нибудь другой словоформе парадигмы, а под флексией — та, которая будет меняться. При этом не предполагается, что основа или флексия является набором подряд идущих символов: в арабской словоформе *katabtu* «я написал» основой считаются буквы *k-t-b*, а флексией, соотвественно, *a-a-tu*.

Естественно, в действительности часто случается, что лексема не обладает одной основой, одинаковой во всех формах, — например, в случае супплетивизма (*иду — шёл*) или регулярного чередования (*лететь — лечу*). Однако и в таком случае обычно всё же удаётся разделить словоформу на основу и флексию, выделив несколько алломорфов основы. При этом каждая флексия с каким-то (или какими-то) из алломорфов может употребляться, а с какими-то — нет. Чтобы описать лексику и словоизменение языка, нужно перечислить все алломорфы основы всех лескем, все возможные флексии и указать, какой алломорф основы с какими флексиями сочетается, т. е. может образовать словоформу.

Все морфологические единицы (основы и флексии) в формате UniParser задаются строками, в которых могут использоваться буквы и ряд специальных символов. Для описания основ и флексий наиболее важны буквы и точка. Буквы имеют своё обычное значение, а точка означает место возможного присоединения другой морфологической единицы: точка в основе обозначает место присоединения флексии и наоборот. В частности, если точка находится внутри морфологической единицы, это означает, что она разрывна. Чтобы образовать словоформу из основы и флексии, нужно совместить их, поставив на место точек в основе соответствующие части флексии (или наоборот).

Для примера возьмём русскую словоформу *краю*, в которой выделяется основа «*кра.*» и флексия «*.ю*»[1]. Наложение основы и флексии, где точке в конце

1 Подчеркнём, что речь идёт о «графических» основе и флексии — сегментах, на которые удобно разбить словоформу для описания словоизменения, но которые в отдельных случаях могут не иметь ничего общего с лингвистически корректным делением словоформ на морфологические единицы.

основы соответствует флексия, а точке в начале флексии, соответственно, основа, даёт словоформу *краю*:

к р а	.
.	ю

В качестве более сложного примера рассмотрим арабскую словоформу *katabtu*. Её основу можно записать как «*.k.t.b.*» — типичный трёхбуквенный семитский корень, а флексию — как «*.a.a.tu*». Точки в основе означают, что и слева от основы, и справа, и в любом месте внутри может находиться часть флексии. И основа, и флексия содержат по три непустых сегмента. Чтобы соединить их в одну словоформу, нужно совместить их, собирая сегменты основы и флексии по порядку, начиная с первой части основы, т. е. действуя по следующему алгоритму:

1. В начале основы стоит точка, поэтому переходим к флексии.

2. Берём первый сегмент флексии — все символы от начала флексии до первой точки. Поскольку он пуст, не производим никаких действий и переходим к основе.

3. Берём первый сегмент основы — *k*, добавляем его к словоформе; видим точку, переходим к флексии.

4. Берём сегмент *a*, добавляем его к словоформе; видим точку, переходим к основе.

5. ...

Проиллюстрируем этот процесс совмещения таблицей:

.	k	.	t	.	b	.
	.	a	.	a	.	tu

Однако во многих языках — например, в большинстве европейских языков, большинство основ и флексий выглядит намного проще: основа имеет единственную точку в конце, означающую, что флексия целиком присоединяется справа от неё, а флексия имеет единственную точку в начале, показывающую, что основа находится слева от неё. Именно так устроена

лексема, разобранная в первом примере.

Основы лексем и флексии перечисляются в разных файлах: stems.txt и paradigms.txt. Флексии в файле paradigms.txt объединены в парадигмы. В каждой парадигме перечисляются флексии одного словоизменительного типа, т. е. показатели, позволяющие получить все формы любой лексемы, принадлежащей к этому словоизменительному типу. Каждый такой словоизменительный тип имеет название; в файле с лексемами указывается, к какому словоизменительному типу относится каждая лексема. Такая организация данных имеет параллели с грамматическим словарём Зализняка: файл stems.txt соответствует основному содержанию словаря, где указываются слова со ссылками на словизменительные типы, а файл paradigms.txt — расположенным в вводной части словаря словоизменительным таблицам.

Не вдаваясь в подробности, мы приведём здесь пример описания лексемы из осетинского словаря и фрагменты парадигмы того словоизменительного типа, к которому она принаждежит:

```
-lexeme
 lex: арц
 stem: арц.|арц.|æрц.
 paradigm: N1
 gramm: N,inanim,nonhum
```

Эта лексема имеет три основы, каждая из которых употребляется с некотоым подмножеством флексий из парадигмы N1. При каждой флексии парадигмы указано, какой вариант основы с ней может использоваться. Ниже приводится фрагмент парадигмы N1:

```
-paradigm: N1
 -flex: <1>.ы
```

```
 gramm: sg,gen
 gloss: GEN
-flex: <0>.æн
 gramm: sg,dat
 gloss: DAT
-flex: <2>.ттæ
 gramm: pl,nom
 gloss: PL
-flex: <2>.тт|ы
 gramm: pl,gen
 gloss: PL|GEN
```

Указание глосс и разбиения на морфемы не является обязательным и необходимо только в том случае, когда создателям корпуса нужно получить глоссированные тексты.

Уже с помощью самых базовых средств, описанных выше, формат UniParser позволяет одинаково легко описывать словоизменение языков разной морфологической структруры: суффиксы, префиксы, инфиксы, полиаффиксы описываются в этом формате с помощью одинаковых средств.

Кроме приведённых выше средств для описания лексики и словоизменения языка в файлах stem.txt и paradigms.txt, в формате UniParser предусмотрен ряд других конструкций для описания различных морфологических явлений.

Регулярные чередования в основе могут быть описаны не только полным перечислением основ. Одним из альтернативных способов это сделать является возможность записать часть основы в флексию, указав, что при глоссировании эта часть не должна считаться частью флексии (этот способ удобен, например, при описании беглых гласных в основе, которые имеются в одних формах, но отсутствуют в других). Другим способом является задание правила, автоматичекси порождающего все или часть основ из одной с помощью языка

регулярных выражений. Ссылка на такое правило может быть указана при парадигме какого-либо словоизменительного типа, в результате чего оно будет применено автоматически ко всем лексемам этого типа. Такой способ является более сложным, однако он позволяет описать любое регулярное изменение в основе.

Несколько разделов посвящено делению на морфемы и глоссированию. В простом случае, когда аффикс разбивается на последовательно соединённые друг с другом морфемы, глоссы приписываются им так, как в примере выше (флексия <2>.тт|ы). Однако имеется несколько сложных случаев: разрывные морфемы, которым должна соответствовать одна глосса, возможность маркирования нулевого показателя и другие. В соответствующих разделах главы 2 предлагаются решения этих проблем.

Для решения проблемы агглютинативных языков, для которых перечисление всех комбинаций аффиксов было бы слишком трудоёмкой задачей, предусмотрено разбиение парадигмы на несколько подпарадигм. Например, если в языке к основе существительного могут последовательно присоединяться показатель числа, показатель падежа и показатель посессивности, формат UniParser даёт возможность описать отдельно эти три парадигмы (т. е. наборы числовых, падежных и посессивных показателей, включая нулевые) и поставить ссылки с одной на другую, указав тем самым, что после любого числового показателя в словоформе должен следовать падежный, а после любого падежного — посессивный. Для описания сочетаемости аффиксов или возможности отсутствия каких-то аффиксов в цепочке предлагаются специальные средства.

В формате также имеются средства для описания редупликации (с помощью регулярных выражений), вариативности основ и аффиксов, форм-исключений и дополнительных сведений об элементах словаря (например, переводов лексем на другие языки).

Описание продуктивного словообразования

Под продуктивной словообразовательной моделью в формате UniParser понимается правило, позволяющее для каждой лексемы *x* из некоторого «естественного множества» *X* регулярным образом получить другую лексему *x'*. Под естественным множеством лексем здесь понимается множество, которое можно задать небольшим количеством простых грамматических или фонетических критериев, например, «все переходные глаголы» или «все существительные с основой, заканчивающейся на согласный».

Задание продуктивных словообразовательных моделей может существенно упростить работу по наполнению словаря, поскольку при наличии такой модели словарь автоматически пополняется лексемами, образованными с её помощью. Поэтому формат UniParser предусматривает набор средств для описания регулярного словообразования. Вся информация о словообразовательных моделях содержится в файле derivations.txt.

В соответствующем разделе главы 2 приводится обзор дискуссии о проблеме противопоставления словоизменения и словообразования. Решение о том, описывать ли данное явление как словоизменительное или словообразовательное, при создании корпусов приходится принимать довольно часто. В связи с этим в исследовании предлагаются практические криетрии для выбора способа описания (экономия усилий при описаний и предполагаемые поисковые запросы к корпусу) и их следствия.

Базовым элементом файла derivations.txt является описание одной продуктивной словообразовательной модели (деривации). В описании этой модели указывается, какие изменения необходимо внести в свойства исходной (деривируемой) лексемы, чтобы получить из неё деривированную, и может указываться, к каким лексемам эта деривация применима. Описание деривации выглядит в целом так же, как описание лексемы, за тем исключением, что вместо перечисления свойств — основы, грамматических значений и т. п., в деривации перечисляются правила, позволяющие получить значения этих свойств для деривированной лексемы.

Ниже приведён пример из осетинского языка — образование перфективной формы глагола при помощи преверба *ных-*:

```
-deriv-type: V-nyx
 lex: <0>ных[.]ын
 stem: ных[.].
 regex-stem: х[^ъ].*
 gramm: +pv,pv-ny
```

Грамматические пометы при применении подобных правил могут добавляться к пометам деривируемых лексем (как в данном случае) или заменять их. Ссылка на деривационное правило может быть указана при конкретной лексеме или в парадигме; в последнем случае словарь будет пополнен деривированными формами всех лексем данного словоизменительного типа.

Важной проблемой при описании словообразования является возможность наследования дериваций, т. е. применения нескольких последовательных дериваций. По умолчанию деривированная лексема не наследует дериваций, ссылки на которые имеются у исходной лексемы. Такому решению есть несколько причин. Во-первых, если, например, у глагола «плавать» заданы деривации, делающие из него причастие «плавающий» и существительное «плавание», то у причастия не должно быть деривации, делающей из него существительное, а у существительного — деривации, делающей из него причастие. Во-вторых, если бы деривации наследовались, то при росте количества дериваций общее количество производных лексем росло бы как факториал, причём среди них было бы много повторов (некторые деривации коммутативны).

Тем не менее, очевидно, что некоторые деривации должны наследоваться деривированными лексемами. В формате UniParser предусмотрены

специальные средства для описания правил наследования. Чтобы производные лексемы могли наследовать некоторые деривации, при задании ссылки на деривацию используется числовой параметр `recurs_class`, по умолчанию равный нулю. Производные лексемы, полученные с помощью дериваций, у которых `recurs_class = n`, наследуют те и только те деривации, у которых `recurs_class < n`. Если ссылка на деривацию d_1 наследуется лексемой, деривируемой при помощи деривации d_2 и при ссылке на деривацию d_1 в поле `stem` явно указана основа, значение этого поля в деривированной лексеме претерпевает те же изменения, что и основа самой лексемы.

С помощью похожих механизмов в формате UniParser описываются регулярное словосложение и инкорпорация.

Описание аналитических конструкций

Хотя разметка аналитических конструкций не являлась первоочередной целью (и далеко не всегда может быть выполнена автоматически с нужной степенью точности), в формате UniParser предусмотрена возможность их описания. Этим средством имеет смысл пользоваться только в тех случаях, когда элементы конструкции довольно жёстко связаны друг с другом (например, при описании греческого кондиционалиса / будущего времени, где частица и глагольная форма, входящая в конструкцию, могут быть разделены не более чем 3 словами из закрытого списка).

В заключительных разделах шдавы 2 приводится обзор дополнительных файлов, с помощью которых можно задать равноценные варианты написания (например, указать, что ё в русских текстах может быть заменено на е), знаки препинания, используемые в данном языке и символы, которые при анализе следует игнорировать.

Глава 3 посвящена обоснованию применимости формата UniParser для решения широкого спектра задач при разметке текстов на разноструктурных языках. В ней представлен ряд нетривиальных морфологических явлений из

конкретных языков и показано, как эти явления могут быть описаны в рамках формата UniParser. Эта глава состоит из следующих разделов.

1. Сингармонизм. На примере калмыцкого языка показано, как можно описать выбор одного из нескольких показателей, различающихся рядом гласного, без явного указания типа основы при каждой лексеме. На примере венгерского языка показано, что такой способ не всегда является оптимальным, и в некоторых случаях более рациональным подходом является указание типа основ.

2. Агглютинация. В этом разделе рассматривается луговой марийский язык. Этот язык является агглютинирующим, при этом именное словоизменение осложнено тем, что показатели числа, падежа и посессивности могут присоединяться к основе в разном порядке; конкрентный набор разных вариантов порядка следования зависит от падежа. Эта проблема успешно решается с помощью задания нескольких подпарадигм с дополнительными ограничениями.

3. Полисинтетизм и «грамматика порядков». На примере адыгейского языка показывается, как формат UniParser может быть использован для описания языка, имеющего множество словоизменительных префиксов и суффиксов, занимающих определённые позиции относительно основы. Кроме того, в этом пункте обсуждается проблема невозможности различить при поиске наборы граммем, отличающиеся лишь их порядком следования из-за того, что набор граммем обычно рассматривается как «мешок свойств».

4. Трансфиксы. В этом разделе в качестве примера описания трансфикса приведено описание арчинского трансфикса имперфектива *r-r*.

5. Метатеза. На примере грузинского масдара показано, как с помощью формата UniParser может быть описана регулярная метатеза, т. е. перестановка фонем в одном из алломорфов основы по сравнению с другим.

6. Редупликация. В качестве примера на редупликацию рассматриваются категории залога и фокуса тагальского глагола, где в глагольной форме

одновременно могут присутствовать редупликация и инфикс, вставляемый после первого согласного основы. Сложность и необычность этого примера в том, что при комбинации этих двух явлений показатель фокуса инфигируется после первого согласного редуплицированного фрагмента основы.

7. Продуктивное словообразование. На примере языка командорских алеутов показано, как может быть использовано средство описания продуктивного словообразования в случае, когда в языке имеется большое количество деривационных моделей, которые могут применяться последовательно в длинной цепочке. В языке командроских алеутов взаимное расположение деривационных аффиксов может быть описано в терминах грамматики порядков: за основой следует 9 слотов, которые могут заполняться теми или иными словообразовательными показателями. В примере приводится разбор этого случая, показывающий, какие из аффиксов следует описывать при разметке как продуктивные словообразовательные модели, и приводится описание наследования этих моделей с помощью свойства `recurs-class`.

8. Инкорпорация. В качестве примера инкорпорации приведена инкорпорация актантов в алюторском языке.

9. Тоновые морфемы. В этом разделе приведён пример описания «тоновых морфем», т. е. аффиксов тонового языка, присоединение которых к основе меняет все или некоторые из её тонов, или «морфем-операций», действие которых состоит исключительно в изменении тонов основы. В качестве примера такого языка рассматривается один из диалектов языка шона (группа банту). Варианты основы с разным набором тонов записываются как разные алломорфы, при этом показывается, как одну из форм можно автоматически получить из другой с помощью правил, описанных в формате UniParser.

10. Аналитическая парадигма. В некоторых случаях имеется необходимость размечать не только морфологически выражаемые категории, но и категории, выражаемы аналитически. В данном разделе рассмотрен крайний

пример этого типа — «аналитическая парадигма» полинезийского языка пилени, в котором нет морфологически выражаемых глагольных категорий, но есть ряд тесно связанных с глаголом проклитик, которые выражают TAM-значени, почти во всех случаях взаимоисключительны и фактически находятся в парадигматических отношениях друг с другом. В примере разбираются все эти проклитики, указывается, какие из них можно было бы считать выражающими словоизменительные значения, и показывается, как это можно сделать в формате UniParser с помощью средства для описания аналитических конструкций.

В **главе 4** приводится краткий обзор технических деталей реализации парсера, работающего с форматом UniParser, и характеристика корпусов, созданных в 2011 г. в рамках программы фундаментальных исследований Президиума РАН «Корпусная лингвистика», а также перспективы дальнейшей работы.

В ходе работы автором был создан парсер, работающий с большей частью средств формата UniParser. Основная часть парсера написана на C++ и использует хэш-таблицы в качестве основной структуры данных для анализа словоформ. Перед анализом файлы с описанием языка проходят предварительную обработку, в ходе которой собираются полные парадигмы, вычисляются все варианты основ и вся информация записывается в специальный файл в формате csv. При разборе парсер получает на вход этот csv-файл с данными о языке и множество чистых неразмеченных текстов в кодировке UTF-8. На выходе размеченные тексты представляют собой XML-файлы, формат которых похож на используемый в Национальном корпусе русского языка. Ниже приводится пример размеченной словоформы с двумя альтернативными разборами из албанского корпуса:

```
<w><ana lex="sy" gr="S,m,inanim,def,pl,acc"
```

```
transl_en="eye"></ana><ana lex="sy"
gr="S,m,inanim,def,pl,nom"
transl_en="eye"></ana>sytë</w>
```

В 2011 г. были созданы грамматические описания и словари (или фрагменты словарей) албанского, калмыцкого, лезгинского и осетинского языков. Эти файлы и созданный автором парсер были использованы для создания соответствующих корпусов; в 2012 г. продолжилась работа над этими корпусами, а также началась работа над корпусами новогреческого и бурятского языков (в работе над составлением словарей и грамматического описания новогреческого языка в формате UniParser автор принимает личное участие).

В **заключении** приведены основные итоги работы:

1. Были изучены типичные проблемы и задачи, связанные с обработкой большого количества текстов для создания корпусов с морфологической разметкой.
2. Был выработан ряд критериев, которым должны удовлетворять морфологический парсер и формат описания языка, используемые для создания таких корпусов.
3. Был разработан формат описания языка UniParser, который по релевантным в исследуемой ситуации параметрам превосходит существующие аналоги; на ряде конкретно-языковых примеров было продемонстрировано, что данный формат подходит для описания широкого спектра морфологических явлений из разноструктурных языков.
4. Был создан пилотный вариант парсера, работающий с большей частью элементов формата UniParser.
5. С помощью разработанных формата и парсера были созданы корпуса албанского, калмыцкого, лезгинского и осетинского языков и создаются в настоящее время корпуса новогреческого и бурятского языков.

Содержание диссертационной работы отражено в следующих **публикациях**:

1. Архангельский Т. А. Электронные корпуса албанского, калмыцкого, лезгинского и осетинского языков // Научно-техническая информация. Серия 2: Информационные процессы и системы, 2012. № 4. С. 24—29

2. Архангельский Т. А., Панов В. А. Аспект в греческом языке: проблемные зоны и типология // Acta linguistica Petropolitana. Труды Института лингвистических исследований РАН, том VIII, часть 2. Исследования по теории грамматики. Вып. 6: Типология аспектуальных систем и категорий. СПб: Наука, 2012. С. 122—148

3. T. Arkhangelskiy, O. Belyaev. A Comparison of Eastern Armenian and Iron Ossetic Spatial Systems // Languages and Cultures in the Caucasus. Papers from the International Conference “Current Advances in Caucasian Studies”, Macerata, January 21–23, 2010. München – Berlin: Verlag Otto Sagner, 2011, pp. 285—299

4. Т. Архангельский, М. Даниэль, М. Морозова, А Русаков. Korpusi i gjuhës shqipe: drejtimet kryesore të punës [Корпус албанского языка: основные направления работы] // Shqipja dhe gjuhët e Ballkanit — Albanian and Balkan Languages. Konferencë e mbajtur në 10-11 dhjetor 2011, Prishtinë [Албанский язык и балканские языки. Конференция, состоявшаяся 10-11 ноября 2011 года в Приштине]. Приштина: ASHAK (Академия наук и искусств Косово), 2012. С. 635—642

5. Сичинава Д. В., Архангельский Т. А. Параллельные белорусско-русский и русско-белорусский корпусы: совместный проект национального корпуса русского языка // Труды Казанской школы по компьютерной и когнитивной лингвитике TEL-2012. Казань: Изд-во «Фэн» Академии наук РТ, 2012.С. 54—60.