АКТУАЛЬНЫЕ ЗАДАЧИ МОРФОЛОГИЧЕСКОГО АНАЛИЗА И СИНТЕЗА В ИНТЕГРИРОВАННОЙ ИНФОРМАЦИОННОЙ СРЕДЕ STARLING

С. А. Крылов (krylov@rinet.ru)
(Москва, Институт востоковедения РАН)
С. А. Старостин (starling@rinet.ru)
(Москва, РГГУ, Центр компаративистики)

Сегодня система STARLING способна строить полные акцентуированные парадигмы всех лексем, представленных в “Грамматическом словаре” А. А. Зализняка (“ГС”), а также осуществлять автоматический морфологический анализ русских словоформ в том же объёме.

Анализ осуществляется не только в виртуальном, но и в актуальном режиме: используя функцию GRAMMAR (V), текст любого размера можно преобразовать в его полный морфологический разбор.

Перечислим задачи, которые в настоящее время система STARLING не умеет решать, но которые в перспективе решаемы без кардинальной перестройки основных принципов её построения.

I. Проблемы графематического анализа

Морфологическому анализу должен предшествовать анализ **графематический**. В его предмет должны войти не только **алфавитные графемы (“буквы”)**, но и внеалфавитные.

1. Внеалфавитные (небуквенные) графемы делятся на сегментные и несегментные.

К **сегментным** внеалфавитным графемам относятся знаки препинания и знаки “внутрисловных” границ (“межморфемный” дефис, “межслоговой” дефис, апостроф).

Как знак препинания целесообразно трактовать **“межсловный” дефис**. Трактовка дефиса как буквы облегчает анализ графических слов с дефисными написаниями частей - таких, как *кто-то*, *где-нибудь* и т.п. Однако достижение этого незначительного упрощения оборачивается радикальным усложнением анализа большого числа свободно образуемых **квазикомпозитов** типа *осетин-извозчик*. Если такой квазикомпозит не описан в “ГС”, он не поддаётся анализу. А при трактовке дефиса как знака препинания все такие образования анализируются в той мере, в какой образующие их части являются анализируемыми сами по себе.

Несегментные графемы подразделяются на супрасегментные и “непечатаемые”.

К **супрасегментным** графемам относятся акцентные знаки главного и второстепенного ударений, при печати “накладываемые” на символ гласной буквы, графема заглавности, а также **каллиграфемы** (жирность, курсивность, подчёркивание и сжатие). На акцентные знаки и на графему заглавности следует обратить особое внимание.

Сегодня члены минимальных графических пар, отличающихся местом акцентного знака - таких, как *замок/замок, разрезать/разрезать, дома/дома, воды/водб, стоит/стоіт,* - получают одинаковую трактовку, как омографы (или графические омонимы).

В режиме “**акцентно-ориентированного анализа**” словоформы, являющиеся членами минимальных пар по ударению, должны будут получить разную трактовку.

Графема **заглавности** сегодня игнорируется. Члены минимальных пар, отличающиеся на графему заглавности («парографы», или «графические паронимы»), получают одинаковую трактовку. Настоящими (реальными) полными омографами эти слова становятся лишь в особых условиях:

(а) при **“сплошной капитализации”**, применяемой нередко лишь в составе заголовков книг и статей; такими окказиональными омографами становятся члены пар типа *этап/ЭТАП*, *ран/РАН*, *газ/ГАЗ*, *арап/АРАП* и т. п.;

(б) при **“начальной капитализации”**, применяемой в начале предложения (такими окказиональными омографами становятся *орёл/Орёл*, *роман/Роман*, *роза/Роза*, *вера/Вера* и т. п.).

Разумно отделять супрасегментную графему заглавности от того слова, на буквенный состав которого она оказалась наложенной, и рассматривать её как автономный значимый сигнал. Желательно, чтобы не возникал эффект графической омонимии там, где её реально нет в тексте, и усматривалась бы графическая омонимия там, где она реально есть.

“Непечатаемые” графемы подразделяются на **положительные (протяжённые)** (“пробел”=“раздельность”, “новая строка”, “красная строка”, “абзацный отступ”, “конец стихотворной строфы”) и **нулевые (непротяжённые)** (“слитность”). Обычно непечатаемые графемы совмещают **делимитативную (разграничительную) функцию** с **конкатенативной (соединительной)**. Однако по функции в речи они неравноправны. У нулевых графем соединительная функция преобладает над разграничительной, и они в большей мере функционируют в роли **конкатенаторов** (такова графема “слитность”). У положительных (протяжённых) графем , наоборот, разграничительная функция превалирует над соединительной, поэтому они в большей мере функционируют в роли **делимитаторов** (таковы графемы “пробел”, “новая строка” и т. п.).

Традиционно из всех непечатаемых пограничных графем анализатор учитывает лишь пробелы, интерпретируя их несколько упрощённо как чисто разграничительные сигналы, причём единственные разграничительные сигналы. Но такое представление нуждается в пересмотре. Оно представляет собой источник препятствий при анализе таких речевых отрезков, в которых представлены хоть какие-либо отклонения от идеальной одно-однозначной схемы соответствия между формой и функцией графематических сигналов.

1. В словоразграничительной функции иногда выступает отнюдь не пробел, а **дефис**. А именно, это касается образований типа *языку-источнику*, *словаря-справочника* и т.п.

2. В словоразграничительной функции иногда выступает и нулевая графема **слитности**. Это характерно для языков с **полисинтетической морфологией**.

Но обыкновенные синтетические языки (русский, немецкий) обладают также элементами полисинтетизма - полиситетическими формами слов.

Так, при образовании графических композитов (**полисинтетических форм**) типа *стадвадцативосьмиэтажный*, нем. *zweihundertsechsundfünfzig*, графема слитности имеет наряду с соединительной функцией ещё и разграничительную.

По аналогии с анализом древнерусского текста (без пробелов), возникает проблема: как “поймать” этот невидимый делимитатор? Обычно предпочтительна стратегия интерпретации, актуализирующая гипотезу о наличии нулевого (невидимого) делимитатора не “на каждом шагу анализа”, а лишь после извлечения вывода о безуспешности анализа таких графических словоформ, которые не поддаются интерпретации в рамках традиционного морфологического анализа (т. е. анализа “синтетических” форм). Так, например, нет никакой надобности при анализе интерпретировать сегмент *зернохранилище* как представитель полисинтетической формы, реализующей сочетание лексемы *зерно* с лексемой *хранилище*, так как соответствующая сложная основа и без того хранится в словаре. Однако при анализе графических сегментов типа *агропромышленный*, *биотехнология*, *самофинансирование* и т. п. “традиционный” морфологический анализ оказывается безуспешным. Здесь уместен анализ в рамках подсистемы “полисинтетических” форм.

3. графема **“пробел”** (**“раздельность”**) отнюдь не всегда выступает как чистый делимитатор. У неё разграничительная функция отнюдь не всегда превалирует над соединительной.

При образовании устойчивых словосочетаний (речений, фразеологизмов) у неё доминирует функция конкатенатора. Можно относиться к речениям, как к “составным словам” (по И. Е. Аничкову) или “составным лексемам” (по Ю. С. Маслову). Составные лексемы содержат графему “пробел” (“раздельность”).

Следует разграничить сами этапы постепенного приближения к такому описанию речений. Можно опираться на общие соображения, лежащие в области “здравого смысла”.

(А) “n-словные” речения перечислить и задать несколько проще, чем “(n+1)-словные”.

Сначала решается задача поиска двусловных речений, потом идут остальные в порядке возрастания структурной сложности.

(Б) **неизменяемые** речения описать и найти в тексте гораздо проще, чем **изменяемые**. Ибо многие изменяемые речения являются раздельнооформленными.

Но некоторый опыт обращения с раздельноооформленными лексемами есть и сегодня: это составные лексемы, пишущиеся слитно: ср. *пятьюдесятью*, *двумстам*. Соответствующий модуль в системе уже есть, но только работает он с крайне незначительной частью русского словарного фонда. Остаётся расширить круг единиц, обрабатываемых этим модулем.

(В) **Исходная (словарная**) форма речения распознаётся в тексте гораздо проще, чем неисходная (несловарная).

(Г) **“Неразрываемые”** речения гораздо легче поддаются анализу, чем **“разрываемые”**.

(Д) **“Неинвертируемые”** речения гораздо проще поддаются распознаванию в тексте, чем **“инвертируемые”.**

II. Проблемы анализа агглютинативной морфологии

Наряду с флективными механизмами в морфологическом строе русского языка, в нём представлены некоторые агглютинативные подсистемы:

“аттенуатив” (***по**больше, **по**меньше, **по**хуже, **по**лучше, **по**дальше, **по**ближе*...);

“негатив” (***не**формальный, **не**решённый, **не**традиционный, **не**велик, **не**дёшево*...).

Н. В. Перцов, обоснованно считает полезным усматривать в русском языке некоторые особые грамматические значения, выражаемые агглютинативно:

“побуждение к совместному действию”, выражаемое суффиксом *-те* в словоформах *пойдёмте, узнаемте, скажемте...* (с. 153);

“фамильярное побуждение”, выражаемое постфиксом *-ка* в формах *иди-ка, идите-ка, идём-ка, идёмте-ка...* (с. 154-160).

Полезно достроить анализатор правилами анализа агглютинативных форм.

III. Проблема интерпретации аналитической морфологии в русском языке.

В русской морфологии есть некоторые аналитические формы.

Согласно отечественной лингвистической традиции, аналитическими формами признаются лишь формы глагола, выражающие время, наклонение, вид и залог; многие авторы склонны трактовать как аналитическую форму существительного в артиклевых языках сочетание существительного с артиклем.

Весьма обычной является трактовка слов *более* и *самый* как аналитических выразителей сравнительной и превосходной степени прилагательного.

Представляется уместным ещё более широкая трактовка аналитических форм, при которой выразителями аналитических форм являются предлоги и послелоги. Применительно к русскому и другим европейским языкам такую точку зрения обоснованно выдвигал Ю. С. Маслов. Согласно Маслову, составные (аналитические) формы лексем включают, в частности, предложно-падежные формы существительных.

Основная трудность: компоненты аналитической формы в реальном тексте дистанцированы. Они располагаются на некотором отдалении друг от друга, причём между компонентами аналитической формы могут быть вставлены знаменательные слова. Но это трудность - лишь при анализе связного текста, содержащего такую форму.

Но и здесь возможно поэтапное решение. Процессор может анализировать и синтезировать аналитическую форму в отрыве от реального контекста (так, как будто бы она была всегда контактной). Тогда проблема дистанцирования компонентов не встаёт.

Надежду на успех вселяет то, что возможности вклинивания между компонентами аналитических форм весьма жёстко ограничены.

IV. Проблемы разграничения “словоизменительной” и “словообразовательной” морфологии

Здесь необходимо переосмысление самой логики рассуждений, лежащих в основе традиционного разграничения основных сфер морфологии. Оно базируется не на одном критерии, а на целом пучке несовпадающих критериев, которые, однако, обнаруживают друг с другом некоторую устойчивую статистическую корреляцию.

Попробуем “развести” используемые критерии и построить разбиение морфологии на основании д в у х дихотомий.

Первую назовём дихотомией “словообразования” vs. “формообразования”.

Формообразованию присущи коррелятивность, композиционность, продуктивность, нефразеологизованность, стандартность, “меньшая вероятность хранения в памяти как единого целого”. **Словообразованию (лексикализации)** присущи: некоррелятивность, некомпозиционность, непродуктивность, фразеологизованность, нестандартность, “большая вероятность хранения в памяти как единого целого”.

Вторую назовём дихотомией “номинации” vs. “словоизменения”. Номинация (но только синтетическая и полисинтетическая) называется также “основообразованием”.

Номинации (в т. ч. основообразованию) присущи: невхождение в состав обязательной категории, отсутствие связи с синтаксисом.

Словоизменению (“инфлексии”) присущи: вхождение в состав обязательной категории, наличие связи с синтаксисом.

В результате наложения двух дихотомий образуются не две, а четыре клетки:

	Словообразование (=лексикализация)	Формообразование
Номинация (в т.ч. основообразование = деривация)	1. Номинационное словообразование	2. Номинационное формообразование
Словоизменение (=“инфлексия”)	3. Словоизменительное словообразование	4. Словоизменительное формообразование

Клетки 1 и 4 едва ли требуют пояснений, поскольку представляют полярные яркие случаи. С трактовкой клеток 2 и 3 дело обстоит сложнее.

К клетке 2 относятся:

а) причастия;

б) уменьшительные существительные;

в) видовые пары с тривиально выводимыми значениями вида;

г) залоги;

(д) плюральные образования.

К клетке 3 относятся:

(а) творительный падеж существительных в качестве словарных наречий (*бегом, босиком, голышом, нагишом, кубарем...*);

(б) второе лицо единственного числа глаголов в качестве словарных междометий (*шутишь!, шалишь!, вишь!, слышь!, врёшь!..*);

(в) мужской род прилагательных в качестве словарных имён лиц (*дежурный, пожарный, полицейский, часовой, постовой, участковый...*);

(г) женский род прилагательных в качестве словарных названий “помещения” (*столовая, прихожая, ванная, уборная, гостиная...*); (*булочная, парикмахерская, пельменная, бильярдная...*)

(д) средний род прилагательных в качестве словарных названий блюд (*мороженое, заливное, жаркое, сладкое, первое, второе, третье, жареное, солёное...*).

Клетка 4 является более или менее описанной.

Клетка 3 является отчасти описанной (для существительных адъективного типа склонения), отчасти неописанной (для междометий глагольного типа спряжения).

Однако степень её описанности напрямую зависит от попадания в «ГС».

Все аналитические формы, подвергшиеся лексикализации и пишущиеся слитно (*всмятку, вприпрыжку...*), распознаются как самостоятельные единицы из «ГС».

Но другие лексикализованные аналитические формы - а именно, пишущиеся раздельно (*в свете, за городом, на руку, по колено, из дому, под Москвой*) не распознаются, так как содержат внутри себя графему пробела.

Целесообразно трактовать раздельно пишущиеся сочетания по аналогии со слитно пишущимися, то есть идентифицировать их с определённым адресом словарной статьи, описывающей лексическую единицу как единое целое, а не как свободное сочетание.

Клетка 2 сегодня описана в той мере, в которой "ГС" трактует образование соответствующей формы как словоизменение: причастия, вид, залог, число и т. п. В других же случаях (напр., отчества, названия жителей, женщин, самок и т. п.) задача пока не решена.

Она может быть решена в той мере, в которой правила носят регулярный характер, а продукт их применения не обнаруживает фразеологичности.

V. Пополнение словаря и решение задач морфологического анализа

Словарь STARLING'а основан на 3-м издании «ГС». Но «ГС» вышел новым, 4-м, изданием (2003), и в нём содержится приложение, содержащее 8000 важнейших **собственных имён**. Эту часть «ГС» полезно воплотить в рамках STARLING'а.

Другое расширение: желателен перенос какого-нибудь **словаря стандартных аббревиатур** русского языка на компьютерные носители и включение в STARLING.

VIII. Общие принципы

Общий принцип работы процессора - установка на 100%-ную адекватность (то есть полноту и безошибочность) в рамках поставленного узкого класса задач (восходящая к «ГС») - должна быть сохранена при любых достройках и усовершенствованиях STARLING.

IX. Синтаксический процессор и направления его усовершенствования

В системе сегодня есть синтаксический анализатор, способный для значительной доли предложений выдвигать правдоподобные гипотезы о структуре непосредственных синтаксических зависимостей между словоформами, входящими в поверхностно-синтаксическую структуру предложения. Анализатор опирается: (а) на результаты морфологического анализа предложения; (б) на ту совокупность сведений о синтаксических свойствах лексем, которая задана в «ГС»: это (б1) указание на частеречную принадлежность; и (б2) указание на подкласс: для существительных это их согласовательные параметры (род, одушевлённость, отчасти число), а для глаголов – переходность.

При работе над достройками синтаксического анализатора осмысленно задаться вопросом: каким способом можно сформулировать ту совокупность задач, решение которых может быть вообще «под силу» анализатору на начальном этапе его усовершенствования? И какова в принципиальном плане сама последовательность шагов, которые предстоит пройти анализатору на следующих этапах его усовершенствования?

IX.1. Создание модуля, способного безошибочно анализировать и синтезировать предложения, длина которых не превышает N словоформ. Минимально N=1.

IX.2. Создание модуля, способного безошибочно анализировать и синтезировать предложения, лексический состав которых не выходит за пределы M первых по частотности словоформ данного (напр., русского) языка. Минимально M=1.

IX.3. Создание модуля, способного безошибочно анализировать и синтезировать словосочетания, длина которых не превышает P словоформ. Минимально P=2.

IX.4. Создание модуля, способного безошибочно анализировать и синтезировать предложения, словосочетания или конструкции [1], составляющие корпус естественноязыковых примеров к первым Q урокам в учебнике A, описывающем данный (напр., русский) язык для изучающих его как неродной на первом этапе обучения.

[1] Под "конструкцией" в данном случае имеется в виду абстрактная единица, реализуемая множеством предложений или словосочетаний, отличающихся друг от друга только лексическим составом.

IX.5. Создание модуля, способного безошибочно анализировать и синтезировать любой связный текст, входящий в «хрестоматию», т.е. R-элементное множество «эталонных» текстов на данном языке. Минимально R=1. Здесь уместно пояснить, что под “эталонными” текстами имеются в виду не самые “типичные”, а скорее самые “ядерные” (т.е. “простейшие”) типы текстов на данном языке. Если удастся довести значение R до нескольких десятков (т.е. включить в хрестоматию «двузначное» число текстов), то желательно иметь в составе такой «хрестоматии» (для русского языка), в частности, тексты:

(а) из стандартного букваря (для первоклассников);

(б) из «Азбуки» Л. Н. Толстого;

(в) из первых нескольких уроков учебников русского языка как неродного.

IX.6. Создание модуля, способного безошибочно анализировать и синтезировать предложения, словосочетания или конструкции, составляющие корпус естественноязыковых примеров в любой «эталонной» («стандартной», «канонической», «авторитетной», «академической») грамматике данного (напр., русского) языка, входящей в заранее заданное множество из S описательных грамматик данного языка (приблизительно упорядоченных “на глазок” по объёму в порядке возрастания общей длины грамматики).

Имеет смысл начинать с таких грамматик, где корпус примеров не является нулевым. Реальные грамматики русского языка, не превышающие по объёму 2 п.л., обычно не содержат примеров вообще, а грамматики объёмом менее 15-20 п.л., не изобилуют примерами; так что сколько-нибудь реально осмысленной задача становится для грамматик, объём которых составляет приблизительно 15-20 п.л. или более.

IX.7. Создание модуля, способного фильтровать результаты работы синтаксического анализатора на основании тех сведений о синтаксических свойствах русских слов (о моделях управления, об ограничениях на синтаксическую позицию слова и т.п.), которые содержатся в компьютерной версии словаря С. И. Ожегова.

IX.8. Эта задача лежит во многом на пересечении задач морфологии и синтаксиса. Создание модуля, способного строить грамматические парадигмы словосочетаний и предложений по аналогии с тем, как устроены морфологические парадигмы синтетических лексем. Ср.:

Gen.Sg.(*каменный мост*) = *каменного моста*;

Abl.Pl. (*каменный мост*) = *от каменных мостов;*

SuperEss. (*каменный мост*) = *на каменном мосту;*

Comit.Sg. (*каменный мост*) = *с каменным мостом;*

Ess. (*Крым*) = *в Крыму;*

Lat (*Крым*) = *в Крым;*

Ess. (*Кавказ*) = *на Кавказе;*

Lat (*Кавказ*) = *на Кавказ;*

Ess. (*там*) = *там;*

Lat (*там*) = *туда;*

Elat (*там*) = *оттуда;*

Dat. (*три сестры*) = *трём сёстрам*;

Praes. (*Петров был дворником*) = *Петров - дворник*;

Praet.Tempor. (*Петров - дворник*) = *Петров был дворником*;

Inf.Fem.Sg. (*красивый*) = *быть красивой*;

Inf.Pl. (*красивый*) = *быть красивыми*;

Approx. (*два часа*) = *часа два*;

Praet.Fem. (*два*) = *было две*;

Fut.Fem. (*два*) = *будет две*;

Gen. (*пить воду*) = *того, чтобы пить воду*;

Dat. (*пить воду*) = *тому, чтобы пить воду;*

Gen. (*мама мыла раму*) = *того, что мама мыла раму*;

Dat. (*мама мыла раму*) = *тому, что мама мыла раму;*

Литература

1. П е р ц о в 2 0 0 1 - Перцов Н. В. Инварианты в русском словоизменении. М.: Языки славянской культуры, 2001. - 279 с.

2. З а л и з н я к 2 0 0 3 - Зализняк А. А. Грамматический словарь русского языка. Издание 4-е, исправленное и дополненное. М.: Русские словари, 2003. - 795 с.