ЕЛЕНА ГРИШИНА, ИЛЬЯ ИТКИН, ОЛЬГА ЛЯШЕВСКАЯ, МАРИЯ ТАГАБИЛЕВА

О задачах и методах словообразовательной разметки в корпусе текстов

1. Введение

Практика создания современных корпусов предполагает в первую очередь разметку данных на уровне слова (например, разметку лемм, частей речи, грамматических признаков и т.д.), а также единиц, более крупных, чем слово (например, разметку синтаксических групп, коммуникативного членения предложения, приведение сведений о тексте в целом и т.д.). Нужна ли в корпусе информация о двусторонних единицах, меньших чем слово, таких, как корни, приставки и суффиксы?

В данной статье мы хотим обрисовать перспективы корпусно-ориентированного подхода к изучению русского словообразования, а также показать, какие практические возможности может предоставить словообразовательная разметка пользователям корпуса.[1]

Существующие описания русского словообразования включают в себя более или менее полные перечни словообразовательных моделей (например, *под – √ – ени – е: вед, ключ, нес, ош, тверждд, чин* [Кузнецова, Ефремова 1986], ср. также [Кубрякова 1965, Townsend 1968, Шанский 1968, Земская 1973, 1992, Развитие.. 1975., Улуханов 1977, 1996, РГ 1980, Шанский, Тихонов 1981, Ефремова 2000]) и списки словообразовательных гнезд (например, *смелый*, *смело, смелость, смельчак, смелеть, осмелеть, осмелиться, посмелеть*; см. словообразовательные словари [Тихонов 1985, Потиха 1961, Тихонов 1978, Wolkonsky, Poltoratzky 1969] и словарь морфем [Кузнецова, Ефремова 1986]). Однако есть проблема, которая, как нам кажется, все еще не нашла удо-

[1] Данная работа выполнена в рамках Программы фундаментальных исследований ОИФН РАН «Текст во взаимодействии с социокультурной средой: уровни историко-литературной и лингвистической интерпретации (2009-2011гг.)»

влетворительного решения – она касается *продуктивности* словообразовательных моделей. В настоящее время под этим термином понимается прежде всего продуктивность в словаре. Продуктивность оценивается относительно всего словарного состава языка (например, непродуктивна модель с суффиксом *б(а)*, ср. *судьба*) или определенного лексического класса (например, модель *с суффиксом ец* непродуктивна для класса слов с адъективным корнем, ср. *глупец*, в то время как модель с суффиксом *ин(а)* продуктивна для класса названий животных, ср. *оленина*). Продуктивность может также оцениваться в диахроническом ключе, но опять-таки сквозь призму словаря (ср. появление суффикса *ирова(ть)* в XVII в. [Изменения... 1964: *44*], вовлечение в словообразовательные модели заимствованных и новых слов).

Вместе с тем продуктивность можно понимать и как вероятность реализации словообразовательной модели в тексте. В самом деле, легко представить себе иностранца, которому непонятен выбор – в конкретном контексте – между диминутивом и недиминутивной формой, между разными моделями образования отглагольного имени и т. д. Не следует также забывать, что для носителя языка словообразование – это живая деятельность, проявляющаяся в речи в виде окказионализмов, языковых игр и т.д.; ср. показательное название книги Е.А. Земской “Словообразование как деятельность” [Земская 1992]. До сих пор продуктивность словообразования в тексте и речи изучалась в основном в стилистическом аспекте [Виноградова 1984 и др.]. Однако, как кажется, любопытно было бы проанализировать, как реализация словообразовательных моделей в тексте связана с реализацией других конструкций; как одни словообразовательные модели сочетаются с другими; как отличается поведение глагольных и именных корней; каким образом однокоренные слова задействованы в установлении кореференции; какова частотность той или иной модели в корпусе в целом или в том или ином жанре, ср. [Пазельская 2009]; также небезынтересно было бы проследить микро-изменения в процессах словообразования (например, какова скорость вовлечения в словообразование новых слов и др.).

Все эти возможности может предоставить словообразовательная разметка корпуса, выполненная с привлечением электронного словообразовательного словаря и снабженная поисковой системой. Данная статья представляет собой проспект проекта, нацеленного на создание полноценного словообразовательного модуля в Национальном корпусе русского языка (http://ruscorpora.ru). Излагаются задачи первого этапа – составления словообразовательной базы данных, ориентированной на разметку корпуса. Поскольку база данных представляет по сути словарь, но реализованный в электронном виде, у нас есть возможность совместить два формата – традиционный словарь морфем и традиционный словообразовательный словарь, то есть можно будет с его помощью выяснить морфемное членение интересующего слова, найти все слова с конкретным корнем (словообразовательное гнездо) или же списки слов с тем или иным аффиксом или сочетанием морфем (словообразовательную модель). Основной акцент при разметке базы данных делается на кодировании плана выражения словообразовательных единиц – их алломорфов, чередований и порядка следования.

Далее в статье мы обсудим общетеоретические проблемы в подходах к словообразованию, без решения которых нельзя обойтись при сплошной разметке словаря (раздел 2); проиллюстрируем предусмотренные поисковые возможности и формат разметки (разделы 3 и 4); очертим план работы и возможные подходы к автоматизации разметки (разделы 5 и 6).

2. Общетеоретические проблемы

Русский язык, обладающий чрезвычайно обширным инвентарем словообразовательных средств и разветвленной системой правил взаимодействия морфологии и фонологии, представляет собой достаточно сложный объект описания с точки зрения словообразования. Существует множество исследований по русскому словообразованию, в том числе и работы по морфотактике (сочетаемости морфем в слове), но единого системного описания русского словообразования (в отличие от словоизменения, описанного А.А. Зализняком в его книге "Русское именное сло-

воизменение" [Зализняк 1967] и Грамматическом словаре [Зализняк 1977/2003]), не существует. Поэтому исследователь, поставивший перед собой практическую задачу составления словообразовательного словаря, неизбежно сталкивается с целым рядом теоретических вопросов, для которых не существует общепринятых ответов и без решения которых осуществление адекватного и последовательного морфемного анализа становится практически невозможным.

К сожалению, существующие словари – словообразовательный словарь [Тихонов 1985] и словарь морфем [Кузнецова, Ефремова 1986] – решают далеко не все возникающие вопросы; более того – зачастую из-за различий в подходах и направлении описания («от слова к слову» или «от слова к морфеме») они дают на них противоречащие друг другу ответы.

Именно поэтому в самом начале нашей работы одной из главных задач стало выявление общетеоретических проблем, которые могут встать перед разметчиком или разработчиком алгоритма автоматической разметки словника, а также выработка системы принципиальных последовательных решений. Поскольку основная цель проекта – в первую очередь практическая (а именно, словообразовательная разметка Национального корпуса русского языка, а не фундаментальное описание системы русского словообразования как таковое), то и общее направление поиска решений было ориентировано на максимальную формализацию и упрощение процесса разметки.

Основные теоретические проблемы, осложняющие процесс морфемного анализа, подробно описаны в [Кузнецова, Ефремова 1986: *3-9*]. К ним относятся в первую очередь проблемы семантики (степень опоры на семантику аффиксов при словообразовательном анализе), омонимия (омоморфия, проблема омонимичных аффиксов), проблема эквивалентных решений («Одно и то же слово в силу многообразия структурно-семантических ассоциаций его с другими словами языка можно соотносить с несколькими мотивирующими словами (основами). Это неизменно приводит к появлению параллельных синхронных разночтений деривационных структур… и морфемного состава слова, особенно его посткорневой части...» [Кузнецова, Ефремова 1986: 6]), вопрос диахронии (степень опоры

на диахронию при морфемном членении словоформы). К сожалению, далеко не все перечисленные проблемы последовательно решены в Словаре морфем. Кроме того, в ходе разработки параметров и формата словообразовательной разметки нам пришлось столкнуться с целым рядом теоретических проблем, эксплицитно в Словаре морфем не обсуждаемых, а именно: инвентарь используемых в разметке терминов – названий классов морфем, статус конкретного морфа в словообразовательной системе языка, переосмысление словообразовательных связей, возможности отображения в разметке алломорфического варьирования.

Первым вопросом стал инвентарь используемых в разметке терминов – названий классов морфем. Наряду с выделяемыми в классических теориях префиксом, корнем, суффиксом, и интерфиксом («соединительной гласной»), в современных теориях присутствуют и так называемые аффиксоиды (префиксоид и суффиксоид – морфемы, способные сочетаться и с бесспорными корнями, и с бесспорными аффиксами. Вопрос об их существовании, в том числе и в русском языке, – один из спорных и активно обсуждаемых морфологами вопросов [Лопатин 2003, Григорян 1981]. Выделять ли аффиксоиды, и если выделять, то на основании каких параметров, какими свойствами должны обладать морфемы, чтобы им должен был быть приписан соответствующий статус, – проблема нерешенная.

Представляется, что, описывая эти единицы в терминах традиционных классов, все префиксоиды следует разделить на префиксы и связанные корни, а суффиксоиды – на суффиксы и связанные корни в зависимости от того, сохраняет ли та или иная единица возможность самостоятельного употребления не в составе сложного слова. Действительно, при внимательном рассмотрении класс аффиксоидов оказывается достаточно неоднородным. Возьмем, например, такие «префиксоиды», как *мега-* и *авиа-*. С одной стороны, обе эти морфемы обладают широкой сочетаемостью и достаточно четко определимым значением, но с другой, *авиа-,* в отличие от *мега-*, может выступать и как самостоятельный корень, например, в словах *авиация*, *авиатор*, да и значение его кажется гораздо более близким к обычному лексическому, нежели к

значению, выражаемому словообразовательными средствами (т.е. грамматическому значению в широком понимании этого термина), что заставляет признать *авиа-* связанным корнем, а слова типа *авиастроительный* – сложными. В то же время морфема *мега-* таких свойств не обнаруживает и может быть без сомнения отнесена к классу префиксов.

Непосредственно с первой описанной проблемой связана другая, а именно – статус того или иного конкретного морфа в словообразовательной системе языка. Для создания словообразовательного словаря и разметки корпуса требуется классификация морфов, то есть нужно иметь точные списки, включающие в себя максимально большое число морфов языка, и знать, какие пометы им присваивать.

Отдельным вопросом является статус морфа не только в общей словообразовательной системе, но и внутри конкретной леммы. Действительно, в ходе исторического развития языка многие суффиксальные, а иногда и префиксальные производные от того или иного корня утрачивают свою непосредственную семантическую связь с ним и перестают ощущаться носителями языка как производные (ср. известную пару *пить* – *пир*). В результате процесса опрощения появляются новые словообразовательные гнезда, уже не ассоциирующиеся с теми, к которым входящие в них слова принадлежали исторически. В русском языке существует огромное количество случаев, когда проследить связь производного с производящей основой без специальных знаний по этимологии очень сложно. Но где проходит граница между «еще однокоренные» и «уже не однокоренные»? Критерий «ощущаемости/неощущаемости связи носителями языка» носит достаточно субъективный характер, тем более что любой эксперимент подобного рода не будет «чистым»: носители, как правило, не членят слова в повседневной речи, и ответы будут зависеть не столько от языкового чутья, сколько от постановки вопроса и – в какой-то степени – от подготовленности носителя в области лингвистики. Разные словари используют разные подходы: словарь Тихонова – строго синхронный, то есть префиксы и аффиксы не выделяются во всех случаях, где связь между производным и производящей основой «не ощущается» (не

очевидна) – точнее говоря, во всех тех ситуациях, в которых автор словаря усматривает (реальную или мнимую) нерегулярность семантических преобразований, словарь А.И. Кузнецовой и Т.Ф. Ефремовой зачастую обращается к диахронии, ориентация на семантику в нем минимальна. Разумеется, это связано не только с установкой автора, но и с направлением работы: словарь Тихонова – словообразовательный, то есть разметка идет «от слова к слову», словарь Кузнецовой и Ефремовой – словарь морфем, то есть разметка идет «от слова к морфеме». Как утверждают сами авторы Словаря морфем, установить границы между синхронным морфемным анализом («…вычленение в слове морфем на основании слов, бытующих в языке сейчас, хотя и в разных его подсистемах…» [Кузнецова, Ефремова 1986: 8]) и анализом историческим («…восстановление такого строения основ, которое было в слове до утраты (иногда сравнительно недавней) производящей основы и которое порою допустимо и в настоящее время, что можно установить, исходя из принципа аналогии…» [там же]) очень трудно. Следуя за авторами Словаря морфем, при проведении морфемного анализа мы отдаем предпочтение формальному, а не семантическому критерию определения строения слов, используя, таким образом, не только синхронный, но в какой-то степени и исторический подход. В пользу такого решения говорит и то соображение, что «стереть» морфемную границу в спорных случаях практически всегда гораздо проще, чем провести, и в связи с этим представляется правильным предоставить более сложный вариант членения слова, который не всегда может быть получен путем интроспекции. Так, пользователь Национального корпуса русского языка, обнаруживший для слова *навзничь* разметку вида *на-вз-ничь* и отсылку к наречию *ниц*, может без всяких дополнительных усилий счесть соответствующие два слова несвязанными и, соответственно, выделение в первом из них корня *-ничь* неоправданным; напротив, при отсутствии в Корпусе указания на такую связь пользователь, предположивший ее существование, вынужден будет обратиться к этимологическим словарям (“Словарь морфем русского языка”, в котором существование такой связи также признается, малоизвестен и труднодоступен).

В ходе исторического развития языка благодаря существенным изменениям в значениях и утрате большого количества слов в словообразовательной системе могут происходить не только процессы опрощения и, наоборот, осложнения основ, но и процесс *переосмысления* словообразовательных связей. Так, в современном языке слово *дотошный*, восходящее к слову *точный* и исторически имеющее корень *точ*, у большого количества носителей ассоциируется в первую очередь со словом *тошнить* и, соответственно, корнем *тошн-*, а слово *столпотворение*, исторически произошедшее от слова *столп* (корень *столп-*), – со словом *толпа* (корень *толп-*).

Кроме того, в словообразовательной системе русского языка существуют так называемые поглощающие суффиксы – сращения суффиксов вида s1s2, которые требуют (для правильного предсказания акцентуации и ряда других свойств производных) разложения основ вида as1s2 на a+s1s2, а не на as1+s2, даже если слово с основой as1 существует, а слово с основой a – нет, ср. [Зализняк 1985: 60-61]. Примером поглощающего суффикса может служить сращенный суффикс *-чат-* в слове *перепончатый*: с точки зрения морфологии в этом слове выделяются два суффикса (*-к-* и *-ат-*): слова **перепона* не существует, – а с точки зрения словообразования – один (*-чат-*): в силу своих акцентных свойств суффикс *-ат-*, в отличие от суффикса *-чат-*, не может сочетаться с приставочными основами. Такие явления наводят на мысль о необходимости введения двойной разметки, которая отражала бы возникающую в подобных случаях реальную неоднозначность.

Самыми сложными для практического решения проблемами при разработке формы и параметров разметки, а также при разработке схемы ее автоматизации стали алломорфическое варьирование и омонимия аффиксов. Эти особенности словообразовательной системы русского языке не позволяют сделать процесс морфемного членения полностью автоматизированным и ставит вопрос о целесообразности их непосредственного отражения в корпусной разметке. С другой стороны, введение разметки такого уровня подробности значительно расширило бы поисковые возможности и вместе с тем – количество теоретических вопросов, которые можно было бы изучать с помощью такого инструмента, как корпус. В

связи с тем, что схема автоматизации находится только в стадии разработки и совершенствования и списки морфов, составленные для упрощения работы программы-разметчика, не являются окончательными, мы решили на первом этапе отказаться от попыток разрешения проблемы омонимии аффиксов и от морфонологического компонента словообразовательной разметки и временно признать каждый алломорф отдельной единицей. Тем не менее, задача сведения алломорфов в морфемы и различения омонимичных морфем представляется одним из самых важных и перспективных направлений дальнейшей работы.

3. Предусмотренные поисковые возможности

Поскольку словообразовательная разметка НКРЯ – первый проект подобного рода в практике аннотации лингвистически ориентированных корпусов, то одной из первоочередных задач стало составление списка возможных поисковых запросов к корпусу со словообразовательной разметкой. Именно от круга задач, которые может поставить перед корпусом пользователь, зависят формат и степень подробности самой разметки.

Представляется, что самым распространенным поисковым запросом должен стать поиск слов, содержащих конкретную морфему (возможно, конкретный алломорф какой-либо морфемы) или некоторое определенное сочетание морфем. Это дало бы пользователю возможность исследовать лексемы, образованные по конкретной словообразовательной модели, анализировать сочетаемость морфем и частотность тех или иных сочетаний, влияние той или иной морфемы на значение содержащего ее слова и особенности его употребления в тексте, свойства слов, принадлежащих к одному словообразовательному гнезду, считать частотность однокоренных слов разных частей речи и решать еще довольно широкий круг теоретических и практических вопросов. В связи с тем, что принадлежность того или иного алломорфа к конкретной морфеме – факт зачастую не очевидный, необходимо предоставить пользователю доступ к списку морфем и их алломорфов: таким образом, можно будет не задавать искомый морф(ему) вручную, а выбирать из представленного списка. К

сожалению, словарь морфем А.И. Кузнецовой и Т.Ф. Ефремовой дает список только корневых и префиксальных алломорфов, сведенных в морфемы, но не дает подобных списков для суффиксов, и сведение суффиксальных алломорфов в морфемы – одна из самых сложных теоретических задач, которые нам еще предстоит решить.

Кроме того, необходимо предусмотреть возможность поиска не только по конкретным значениям параметров, но и по наличию/отсутствию помет того или иного типа, то есть поиск композитов, поиск слов, содержащих один или более аффикс (например, вполне возможным представляется запрос «найти все слова с двумя приставками»), поиск слов со связанными корнями. Естественно, нужно предоставить пользователю корпуса возможность комбинировать поиск обоих типов (это позволит искать, например, все префиксальные (суффиксальные) производные от конкретного корня). Также должна существовать опция комбинирования поиска по словообразовательной разметке с поиском по другим (семантическим и грамматическим) параметрам, предусмотренным разметкой НКРЯ. Именно эти возможные поисковые задачи предопределили принятый нами формат разметки.

4. Формат разметки

Каждому слову приписывается последовательность словообразовательных тегов, отражающих его морфологическое членение. Размечаются все морфемы основы; морфемы, участвующие в словоизменении (например, окончания имен, суффиксы причастий), разметке не подлежат. В разметке используется принцип представления алломорфов в орфографической записи.

Ниже приведен пример словообразовательной разметки слова *переподготовка*:

```
<w>…
<der cl="pref" pl="1" al="пере" mf="пере"></der>
<der cl="pref" pl="2" al="под" mf="под,подо,подъ,пода"></der>
<der cl="root" pl="3" al="готов" mf="готов,готовл,готавл"></der>
<der cl="suf" pl="4" al="к" mf="к,ок,ек,оч,еч" id="3"></der>
переподготовка</w>
```

Словообразовательные теги кодируют следующую информацию:

1. класс единицы (cl): корень (root), префикс (pref), суффикс (suf); дополнительно в эту же зону помещаются сведения о связанных корнях (adsorb), позиции аффикса после финитной части глагола (post), альтернативном статусе разбора (alt) и др.
2. порядок следования единиц (pl) – целое число от 1 до n;
3. алломорф, реализованный в начальной форме слова (al);
4. морфема (mf) – задается списком алломорфов;
5. индекс (id), позволяющий отличить омонимичные морфемы и кодирующий варьирование алломорфов в словоизменительных формах.

Таким образом, в слове *переподготовка* присутствуют префикс, на первом месте (алломорф *пере*, представляющий одноименную морфему); префикс, на втором месте (алломорф *под*, представляющий морфему *под/подо/подъ/пода*), корень, на третьем месте (алломорф *готов*, морфема *готов/готовл/готавл*) и суффикс, на четвертом месте (алломорф *к*, представляющий морфему с вариантами *к/ок/ек/оч/еч* и реализующийся как *ок* в форме родительного падежа множественного числа).

Формат разметки допускает существование альтернативных разборов. Так, представленный ниже пример позволяет реконструировать двойное членение слова *перепончатый* как *пере-пон-ч-ат-ый* и как *пере-пон-чат*-ый (ср. обсуждавшееся выше явление «поглощения» суффиксов, п. 2):

```
<w>…
<der cl="pref" pl="1" al="пере" mf="пере"></der>
<der cl="root" pl="2" al="пон" mf="п,пин,пон,пя"></der>
<der cl="suf" pl="3" al="ч" mf="к,ок,ек,оч,еч" id="3"></der>
<der cl="suf,adsorb,alt" pl="3" al="чат" mf="чат"></der>
<der cl="suf" pl="4" al="ат" mf="ат,ят"></der>
перепончатый</w>
```

5. Порядок разметки

Следующим этапом после разработки параметров разметки и ее формата стала разработка собственно технологии аннотации.

Словообразовательная разметка корпуса предполагает работу с полным словником НКРЯ, за исключением отдельных редких и окказиональных слов. Поскольку словник НКРЯ, который все еще находится в процессе составления, значительно превышает по объему даже словник Грамматического словаря А.А. Зализняка [Зализняк 1977/2003] (около 110 тысяч лемм), то первоначально возникла идея взять за основу большой морфемный или словообразовательный словарь, который затем можно было бы дополнить в части новых слов.

На первый взгляд, хорошим претендентом на эту роль кажется словарь А.Н. Тихонова [Тихонов 1985], который обладает чрезвычайно большим объемом словника (154 000 единиц). Однако этот словарь не дает, на наш взгляд, удовлетворительного морфемного анализа в большинстве спорных и сложных случаев (ср. *сказ-к-а, поезд* и др.), и это делает его использование в качестве основы для разметки малоприемлемым.

“Словарь морфем русского языка” [Кузнецова, Ефремова 1986], включает всего 52000 лемм – то есть, если использовать только этот словарь для разметки текстов, то больше половины слов корпуса не получат в таком случае словообразовательных помет. Вместе с тем, сам словарь устраивает нас в том отношении, что в решении основного круга теоретических вопросов мы следуем за его авторами. В результате было принято решение создать собственный морфемно-словообразовательный словарь корпуса, однако взять Словарь морфем за основу: использовать представленные в нем списки аффиксов и алломорфов, а также привлечь данные словаря для разрешения сложных случаев морфемного членения, в частности, при выделении суффиксов. Вместе с тем, был также составлен список поправок, касающихся некоторых конкретных решений Словаря морфем, которые показались нам неприемлемыми. Например, Словарь морфем выделяет в слове *судья* аффикс *ья* (по нашему мнению, в орфографической записи этот суффикс должен иметь вид *ь*), в словах типа *подготавливать* Словарь морфем выделяет *л* как отдельный суффикс, идущий после корня *готав* (мы же считаем его входящим в состав корня *готавл*, чередующегося с *готов*, ср. пару *подготовить – подготавливать*) и т.д.

Итак, поскольку Словарь морфем покрывает лишь небольшую часть словника НКРЯ, речь фактически идет о самостоятельной разметке нового словаря корпуса силами сотрудников проекта (естественно, с привлечением данных словарей на тех участках, где это возможно). Ручная обработка словника подобного объема с предполагаемой нами степенью подробности представляется задачей трудновыполнимой. В связи с этим единственным возможным решением представляется разработка схемы автоматизации морфемного анализа, которая позволила бы проделать большую часть работы по отделению аффиксов в автоматическом режиме. К сожалению, создать точный автоматический разметчик практически невозможно, и полностью избежать вмешательства исследователя в процесс морфемного анализа не удастся, но все же «ручную» часть можно свести к минимуму, заключающемуся только в проверке результатов работы программы, если осуществить хотя бы первичное деление и приписывание морфов в автоматическом режиме.

По ряду причин наименее проблемной зоной для автоматической разметки оказываются префиксы. Во-первых, префиксальные морфы достаточно легко отделяются даже в орфографической записи, проблема проведения морфемной границы (актуальная, например, для глагола *обрыднуть*) возникает в ничтожно малом количестве случаев. Во-вторых, префиксы обладают в общем случае достаточно широкой сочетаемостью, что значительно упрощает разработку алгоритма по их отделению. Большое преимущество задачи отделения префиксов состоит в том, что для них возможно разработать алгоритм автоматической разметки без опоры на другие аффиксы, чего нельзя осуществить ни для корней, ни для суффиксов. К тому же, алломорфическое варьирование не так распространено в зоне префиксов, как в зоне суффиксов, что значительно упрощает решение проблемы сведения алломорфов в морфемы. Одним из главных факторов, облегчающих выделение префиксов, является то, что префиксы образуют «кластеры» гораздо реже, чем суффиксы, и среди префиксальных кластеров фактически не встречается неделимых (по крайней мере, неделимых в орфографической записи).

Разметка суффиксальной части лемм очевидно должна вызвать большое количество проблем в связи с широко распространенным в этой части алломорфическим варьированием и со сращениями суффиксов. Разметка суффиксов будет проводиться с опорой на Словарь морфем А.И. Кузнецовой и Т.Ф. Ефремовой [Кузнецова, Ефремова 1986], а также на работу И.Б. Иткина [Иткин 2007], содержащую наиболее подробное на сегодняшний день описание алломорфического варьирования всех аффиксальных морфем русского языка, кроме заимствованных. Для разметки слов, не вошедших в Словарь морфем, будут построены алгоритмы-эвристики, учитывающие опыт разборов в Словаре. Отдельную задачу составит составление инвентаря аффиксов в заимствованных словах и разметка заимствованных слов – эта задача Словарем морфем не решается.

Без сомнения, самой сложной проблемой при автоматическом морфемном анализе представляется отождествление корней, так как полных списков словообразовательных гнезд, удовлетворяющих нашим целям, не существует. Отделение префиксов и суффиксов должно значительно облегчить задачу по выделению и – частично – по отождествлению корней. Таким образом, правильным порядком разметки представляется следующий: отделение префиксов, затем суффиксов, отождествление корней.

6. Мы делили *а-пельсин*, или как отделить префиксы

Первой задачей на пути морфемного анализа словника НКРЯ стала задача автоматического отделения префиксов. Эта работа включала в себя несколько теоретических и практических этапов, в том числе: составление списков префиксальных морфов, которые послужили бы основой работы программы-разметчика, разработка схемы автоматизации, написание программы-разметчика в соответствии с разработанной схемой, ручная проверка результатов работы программы, попытка оптимизации работы программы по результатам ручной обработки размеченного словника.

6.1 Вспомогательные списки

Для того чтобы автоматически отделить префиксы и выделить

первые корни сложных слов, необходимо было составить как можно более полные списки префиксальных морфов и связанных частей сложных слов, на которые могла бы опираться программа-разметчик. Несмотря на существующие в разных источниках списки подобного рода, процесс составления полных списков оказался достаточно трудоемким: список префиксальных морфов, данный в Русской Грамматике [РГ 1980], оказался неполным, а список «повторяющихся (в том числе связанных) частей сложных слов» был составлен достаточно непоследовательно. Принцип, по которому авторы выбирали вошедшие в список «повторяющиеся части» из всей массы встречающихся в начале сложных слов, совершенно неясен. Неполнота (с нашей точки зрения) списка префиксальных морфов объясняется общим подходом авторов к морфемному анализу: редкие префиксы, чьи немногочисленные производные претерпели ряд семантических изменений и утратили прозрачную связь с мотивирующей основой (а иногда – и саму мотивирующую основу), авторами [РГ 1980] не выделялись и в список не вошли. Так, например, в Русской Грамматике отсутствует префикс *ку* (*кумекать, скукожиться*), выделяемый, однако, Словарем морфем Кузнецовой и Ефремовой [Кузнецова, Ефремова 1986]. Изъяном же списка префиксов, достаточно последовательно представленного в самом Словаре морфем, является принципиальное отсутствие в нем заимствованных морфов, число которых в русском языке достаточно велико, и морфологический статус которых представляет зачастую отдельную проблему (см. выше о так называемых «аффиксоидах»). Таким образом, списки, ставшие основой работы нашей программы-разметчика, были составлены на основе списков Словаря морфем [Кузнецова, Ефремова 1986] и Русской Грамматики [РГ 1980]. Элементы, вошедшие в список «повторяющихся (в том числе связанных элементов сложных слов)» Русской Грамматики, которые все вместе могли бы претендовать на статус так называемых аффиксоидов, были расклассифицированы на основании их деривационных свойств в две группы: префиксальные морфы и «корни, связанные справа». Таким образом, из используемой нами в разметке системы терминов на основании описанных выше соображений был исключен термин «префиксоид».

6.2 Схема автоматизации разметки

Несмотря на то, что существуют программы, производящие автоматический морфемный анализ слов, программы, отделяющей только префиксы и анализирующей сложные слова, по нашим сведениям не существует. В то же время, разделение этапов морфемного анализа слова представляется достаточно правильным подходом: проверка результатов разбора после отделения морфем одного типа позволяет избежать накопления ошибок, неизбежного при одновременном полном анализе слов: поскольку морфы в слове непосредственно контактируют, проведение одной неверной морфемной границы ведет к приписыванию слову минимум сразу двух неверных помет, то есть ошибка на стадии выделения префикса автоматически влечет за собой неправильное выделение корня.

В связи с этим перед нами встала задача разработки «с нуля» схемы автоматизации отделения префиксальных морфов. Основным свойством словообразовательной системы русского языка, которое позволило достаточно успешно осуществить поставленную задачу, является то, что в большинстве случаев префиксальные производные имеют в языке соответствующую беспрефиксную пару, послужившую для них производящей основой.

6.3 Принцип работы программы

На вход программе подается список лемм (в данном случае – словник Грамматического словаря). Опираясь на списки префиксов и связанных частей сложных слов, программа проверяет наличие соответствующих начальных частей в леммах словника по порядку убывания длины морфа: сначала осуществляется поиск соответствий более длинным приставкам, затем более коротким, с целью уменьшения количества ошибок (например, чтобы в словах с префиксом *между* не был выделен префикс *меж*). Затем программа проверяет наличие в словаре леммы, соответствующей неприкрытой части леммы с выделенным префиксом. Если таковая существует, лемме приписывается свойство «имеет префикс» и указываются префикс и неприкрытая часть основы. Если находятся две или более цепочки букв разной длины (как в случае *меж* и *между*),

соответствующие разным префиксам списка (или префиксу и связанному корню), программа приписывает оба возможных разбора (для разрешения подобных спорных случаев необходима ручная проверка результатов работы программы). Процедура повторяется для всех префиксов списка. После завершение первого круга проверки, программа тем же образом проверяет отделенные неприкрытые части лемм, выделяя таким образом не только первые, но и вторые, и последующие приставки.

Естественно, при использовании описанной выше схемы неизбежным является получение достаточно большого количества неверных разборов, в которых морфемная граница проведена там, где ее на самом деле не существует (ср., например, получившиеся в результате разметки разборы *бес-еда* и *на-гайка*). Достаточное количество подобных разборов также делает необходимой ручную проверку результатов работы программы.

На финальном этапе работы автоматический разметчик проверяет список лемм, не получивших разбора, на наличие связанных корней: ищет одинаково оканчивающиеся леммы, и если начальные части таких лемм являются приставками, входящими в составленный список префиксов, то им приписываются свойства «имеет префикс» и «имеет связанный корень».

Разработанная программа осуществляет также предварительную обработку сложных слов, опираясь на список связанных частей сложных слов (таким образом выделяются сложные слова со связанными начальными корнями) и на следующий принцип: если в списке лемм, не имеющих префиксов, есть леммы, оканчивающиеся на части, идентичные другим леммам из данного списка, а начальные части не соответствуют префиксам и связанным корням, то таким леммам приписывается первичная помета «сложное слово».

Как уже было сказано выше, в связи с предусмотренной возможностью приписывания нескольких разборов и с тем, что принцип работы программы все-таки несовершенен и не дает стопроцентной точности в разборе, результаты работы программы нуждаются в постредактировании, осуществляемом одним или несколькими людьми.

6.4 Результаты работы программы

Разработанная нами по вышеописанному принципу программа работает с точностью, приблизительно равной 90%, что представляет собой достаточно высокую степень точности, учитывая количество нерегулярных случаев словообразования в русском языке. В связи с предусмотренной возможностью нескольких вариантов разборов одной леммы в результате ее работы для 110 000 лексем, входящих в словник НКРЯ, было получено примерно 125 000 возможных разборов. Таким образом, ручная проверка результатов работы программы, как и предполагалось заранее, оказалась неизбежной.

Для оптимизации и ускорения процесса ручной проверки результатов была создана специальная компьютерная программа – рабочее место постредактора. Общий список лемм был разбит на равные части (приблизительно по 20 000 лемм каждая), каждая из которых проверялась отдельно разными участниками проекта. Спорные случаи разбора, а также статус отдельных морфов обсуждались совместно. После первичной проверки отдельные отредактированные части были вновь собраны в единый массив и подвергнуты вторичной проверке на предмет единообразия принятых по спорным случаям решений.

В процессе ручной обработки результатов работы программы (постредактирования) было выявлено несколько проблем. Во-первых, составленные нами списки, служившие основой работы программы, оказались неполными в части «связанных корней сложных слов», что неудивительно, учитывая их количество, а также тот факт, что предварительных полных списков «первых частей сложных слов» у нас в распоряжении не было, и приходилось работать, как описано выше, с достаточно непоследовательно составленным и кратким списком, представленным в [РГ 1980]. С другой стороны, списки префиксальных морфов оказались «слишком полными»: входящие в них редкие префиксы были неверно отделены программой в очень большом количестве случаев (это касается, в первую очередь, префикса *к-*, выделяемого только в нескольких случаях – в наречиях *кверху* и *книзу* и нек. др.; ср. неправильные случаи членения *к-лад, к-рот* и т.п.). Это заставило

нас включить в программу не только списки префиксов, но и списки всех производных для каждого из редких префиксов (а именно, префиксов с не более чем десятью производными), чтобы исключить лишние случаи отделения подобных редких морфем. Кроме того, в результате постредактирования на основании получившихся результатов вручную были пополнены списки «связанных корней сложных слов», что также позволило увеличить процент точности, с которым работает программа.

7. Заключение

Проект рассчитан на три года. Несмотря на то, что значительная часть работы уже проделана (разработаны общая концепция и порядок разметки, а также разработана и написана программа по автоматическому выделению префиксов), в ближайшие два года нам предстоит решить еще довольно большое количество сложных теоретических и практических задач. К сожалению, до сих пор не разработана схема автоматизации разметки суффиксов и корней. К тому же одним из по-прежнему актуальных направлений работы является усовершенствование программы по отделению префиксов с тем, чтобы повысить точность ее работы и свести пост-редактирование («ручную» часть) при разметке добавлений к словнику Грамматического словаря Зализняка (словника НКРЯ) к минимуму.

Одной из самых главных задач как с точки зрения практики (то есть разметки), так и с точки зрения теории станет составление списков морфем с их алломорфами (то есть сведение префиксальных и аффиксальных алломорфов в морфемы) и разрешение омонимии аффиксов (то есть составление списков морфем с различением омонимичных аффиксов). Эта задача представляется очень сложной с теоретической точки зрения и в связи с этим достаточно трудоемкой. Кроме того, благодаря существованию различных подходов к решению вышеописанных теоретических вопросов многие случаи неизбежно должны вызвать большое количество споров. Но, несмотря на все это, подобные списки станут большим шагом на пути к подробному описанию системы слово-образовательных единиц, заполнив важные лакуны в существующих

ные описаниях, таких, как "Словарь морфем русского языка" А.И. Кузнецовой и Т.Ф. Ефремовой [Кузнецова, Ефремова 1986].

Список литературы

Rasch, Barbara J.L. *A syntactic analysis of word-formation in Russian with particular emphasis on deverbal nouns*. PhD dissertation. University of Washington, 1977.

Townsend, Charles E. *Russian word-formation*. Columbus OH: Slavica publishers, 1968.

Wolkonsky C., Poltoratzky M. *Handbook of Russian Roots*. London - New York: Columbia University Press, 1969

Виноградова В.Н. *Стилистический аспект русского словообразования*. М.: Наука, 1984.

Григорян Э.А. *Суффиксоиды в системе современного русского языка (на материале сложений со вторым глагольным компонентом)*. Диссертация на соискание ученой степени кандидата филологических наук. М., 1981.

Ефремова Т.Ф. *Новый толково-словообразовательный словарь русского языка*. М., 2000.

Зализняк А.А. *Русское именное* словоизменение. М., 1967.

Зализняк А.А. *Грамматический словарь* русского языка. М.: Русский язык, 1977. 4-е изд. М.: Русские словари, 2003.

Зализняк А.А. *От* праславянской акцентуации *к русской*. М.: Наука, 1985.

Земская Е.А. *Современный русский язык. Словообразование*. М., 1973.

Земская Е.А. *Словообразование как деятельность*. М., 1992.

Изменения в словообразовании и формах существительного и прилагательного в русском литературном языке XIX века. Очерки по исторической грамматике русского литературного языка XIX века, под ред. В.В.Виноградова и Н.Ю.Шведовой. М., 1964.

Иткин И.Б. *Русская морфонология*. М., Гнозис, 2007

Кубрякова Е.С. *Что такое словообразование*. М., 1965.

Кузнецова А.И., Ефремова Т.Ф. *Словарь морфем русского языка*. М., 1976.

Лопатин В.В. Аффиксоид // *Русский язык. Энциклопедия* / Под ред. Ю.Н. Караулова. М., 2003.

Пазельская А.Г. Модели деривации отглагольных существительных: взгляд из корпуса // *Корпусные исследования по русской грамматике*. М.: Пробел-2000, 2009.

Плунгян В.А. *Общая морфология*: Введение в проблематику. М.: Эдиториал УРСС, 2000

Потиха З.А. *Школьный словообразовательный словарь* (под ред. С.Г. Бархударова). М.: 1961.

Развитие современного русского языка. *1972. Словообразование. Членимость слова*. М., 1975.

РГ 1980: *Русская грамматика*. М.: Наука, 1980.

Тихонов А.Н. *Словообразовательный словарь русского языка*. В двух томах. М.: Русский язык, 1985.

Тихонов А.Н. *Школьный словообразовательный словарь русского языка*. М., 1978.

Улуханов И.С. *Единицы словообразовательной системы русского языка и их лексическая реализация*. М.,1996.

Улуханов И.С. *Словообразовательная семантика в русском языке и принципы ее описания*. М., 1977.

Шанский Н.М. *Очерки по русскому словообразованию*. М., 1968.

Шанский Н.М., Тихонов А.Н. (ред.). *Современный русский язык*: В 3-х ч. Словообразование. Морфология. М., 1981. Ч. 2.

Summary: Word-formation annotation of the Russian National Corpus – aims and methods.

This article reports on a project aimed at creating a full-fledged word-formation annotation of the Russian National Corpus (http: //rus-corpora.ru). In the article we outline the goals of the first stage of our work – compiling a word-formation database suitable for corpus annotation. We furthermore discuss some important theoretical problems of Russian word-formation, describe stipulated searching possibilities and the annotation format, and outline possible approaches to automation of the annotation process.

E-mail: olga.lyashevskaya@uit.no