# Морфанализ in vivo[1]

Т.Ю. Кобзарева (РГГУ, Москва)
stamstam@mtu-net.ru

Работа представляет кратко комментированный обзор проблем автоматического морфологического анализа естественного текста на русском языке – низшего уровня поверхностно-синтаксического анализа. Эти проблемы неизбежно возникают при переходе от словарного морфанализа к собственно синтаксическому. Их решение необходимо для всех последующих этапов поверхностно-синтаксического анализа в рамках любого способа его реализации.

Обсуждаются морфологические и пограничные морфосинтаксические проблемы анализа (1) слов, отсутствующих в словаре, но относящихся к обычным словообразовательным и словоизменительным моделям (продуктивные словообразовательные модели, неологизмы), (2) слов, уже опознанных на этапе словарного морфанализа, но теряющих при вхождении в устойчивые словосочетания морфологическую автономность и морфологические признаки, значимые для синтаксического анализа, (3) несловарных комплексов, выступающих как единое целое и имеющих особые морфосинтаксические свойства: имен собственных, окказиональных аббревиатур и названий в кавычках.

## Введение

Решение любых прикладных задач, связанных с анализом естественного текста, а в нашем случае – с поверхностно-синтаксическим анализом [Кобзарева 1)] – начинается с морфологического анализа. Что ожидается от морфанализа? По-видимому, морфологическая идентификация компонент линейной структуры предложения – определение их морфологического статуса: части речи и соответствующих форм. После появления «Грамматического словаря» Зализняка задачу автоматического морфологического анализа для русского языка можно было бы считать в принципе решенной. Однако универсального модуля морфанализа, доступного всем исследователям, нуждающимся в морфанализе, по-прежнему не существует[2]. С чем это связано? Все проблемы можно условно разделить на две группы: (1) – несовместимость способов представления информации в разных и (2) отсутствие попыток построить единый лингвистический базис для этой области прикладных проблем. Экспериментирование с естественным текстом на любом этапе анализа требует

[1] Работа подготовлена при поддержке гранта РФФИ 03-06-80109.

[2] Коллекция излагаемых проблем - результат построения ПСА [Кобзарева и др. 2001] при работе с «живым» текстом.

решения конкретных проблем данного уровня. Рассмотрим в первом приближении спектр лингвистических задач автоматического морфанализа.

## Проблемы словарного морфанализа

Морфанализ естественного текста на основе словаря оставляет нерешенными множество периферийных проблем[3].

Эти задачи по сложности и объему не сопоставимы, конечно, с объемом задач, решаемых системой словарного морфанализа, но, только решив их, мы получаем возможность работать с естественным текстом, т.е. с **текстом in vivo**.

Необходимой составляющей эксперимента при разработке любой идеальной модели анализа является создание универсальных лингвистических описаний для каждого этапа анализа. Удобно иметь лингвистические описания – коллекции явлений уровней – построенные независимо от используемых в системах форм представления данных.

Так как практически любая задача анализа требует поверхностно-синтаксической компоненты, важно очертить круг рутинных проблем низших этапов поверхностно-синтаксического анализа (ПСА).

Всякая прикладная система ПСА строится на основе представлений авторов об «идеальной структуре» возможного или желаемого результата.

Совершенствуя описания, исследователь для каждой зоны лингвистических явлений постепенно приближается к некоторой «идеальной структуре». **In vitro** мы можем быть близки к цели, но, перейдя к экспериментам с естественным текстом, немедленно сталкиваемся с тем, что все наши ясные **in vitro** построения **в естественных условиях не работают,** так как текст **«засорен»** множеством частных явлений, которые уже на этапе морфанализа выходят за рамки компетенции обычного морфанализа на основе словаря, и вообще – за рамки морфанализа.

В силу того, что жесткой границы между явлениями морфологии и синтаксиса нет, возникает острая необходимость введения отдельного от словарного морфанализа уровня - «постморфанализа» или «морфосинтаксиса», где мы вынуждены решать многочисленные рутинные проблемы очищения текста от всякого рода морфосинтаксического «мусора».

Отличительной чертой прикладных лингвистических задачи является необходимость уметь проинтерпретировать, работая с текстом in vivo, каждое частное нестандартное пограничное явление, не разрушая лингвистической «идеальной картины» модели, заложенной в структурах словарей, правил, алгоритмов, списков, т.е. чтобы интерпретации частных явлений выдерживали проверку при эксперименте in vivo на каждом уровне анализа. И в первую очередь необходимо уметь исчислить множество периферийных, часто рутинных, но безусловно реальных и значимых задач анализа на уровне морфологическом.

После морфанализа на основе словаря мы, помимо морфологических неоднозначностей, возникающих при интерпретации словоформ на основе словаря [Пащенко 1967; Аношкина 2001; Кобзарева 2002, 2)], сталкиваемся со множеством проблем.

---

[3] Речь идет о словарном морфанализ РЯ, первая версия которой была создана Н.А. Еськовой в 70-е годы в Информэлектро в отделе Д.Г.Лахути. Версия была ориентирована скорее на синтез, поэтому в дальнейшем ее пришлось значительно переработать соответственно прикладным задачам анализа.

Все они относятся к явлениям омонимии. Некоторые из них явно относятся к проблемам морфологическим, для других же границу между морфологией и синтаксисом провести трудно, и часто – просто невозможно. Однако хочется повторить, что успешность решения этих по большей мере рутинных проблем в значительной мере определяет возможности системы в целом.

Первая из этих проблем связана с насыщением словаря. Множество слов с обычной парадигмой, почему-либо отсутствуют в словаре. Это или малоупотребительные слова, или неологизмы, архаизмы, диалектизмы, многочисленные продуктивные формы, в словарях не предусмотренные, или другие «почти обычные» слова с нормальной для РЯ парадигмой, характеризующие авторскую индивидуальность.

Для их идентификации в 70-е годы при разработке уже упомянутой системы автоматического морфанализа для ИПС было построено некоторое дополнение к обычному словарному морфанализу.

Этот дополнительный «несловарный» морфанализ (НСМ) на основе словаря концов слов («хвостов»), определяет часть речи и формы слов. Хвост удлиняется до тех пор, пока для некоторой представительной группы словоформ по нему оказывается возможно построить гипотезу о части речи и формах этой группы. Слова-исключения, нарушающие гипотезу, т.е. слова, у которых тем же «хвостам» соответствуют другие грамматические характеристики, вводятся в словарь основ. НСМ описывает словоизменительные типы в объеме [Зализняк 1980].

Такой НСМ для слов с обычным для РЯ словоизменением можно довести до очень высокой степени точности (с учетом омонимии частей речи, которая встречается при этом практически не чаще, чем при словарном морфанализе).

В процессе тестировании модуля сегментации при поиске сложных по сегментной структуре текстов в прозе Набокова, Мандельштама, Л.Н. Толстого, Гоголя часто встречались лексически продуктивные формы, неологизмы, «аномалии», не учтенные, например, в Словаре у А.А. Зализняка и в компьютерном словаре 2000-ого Word-а. Таких слов у каждого автора множество. И наш не чрезмерно тонкий анализ по «хвостам» позволяет строить для них гипотетические морфологические характеристики, обеспечивая тем самым работу модуля сегментации [Кобзарева и др. 2001].

Примеры таких случаев из Набокова и Мандельштама (выделены жирным шрифтом): *... тлена я не боюсь, тем более, что жду такую отраду, такую степень* ***зачеловеческого*** *бытия, которая...; ... падая, хрустя, хохоча с* ***запышкой,*** *влез на сугроб...; ... восклицательные знаки* ***сквозисто*** *горят у него в голове; ... рыбы не* ***уживал...; ...*** *и что* ***близехонько*** *от него; ... засыпая в кровати с* ***ослабнувшей с****еткой...; нелепая, истерическая, суеверная,* ***сверхподозрительная,*** *и чем-то привлекательная мать внушила сыну...; загрязняются непреодолимыми* ***необычностями...; ...*** *все эти создания, чья* ***мельтешня*** *создает; ... я был рад, что в комнате* ***надышано...;*** *... все на какой-то станции* ***повылезли...;*** *Он находился в том состоянии чувств и души, когда существенность, уступая мечтаниям, сливается с ним в неясных видениях* ***первосонья; ...*** *с привычными* ***обобщительными*** *рукоплесканиями...; ... держа... пирамидальную* ***фиоль****, лил ему на макушку..;... и племянник из банка семенил коротенькими ножками и покачивал тяжелой* ***бисмарковской*** *головой; ....сидя в позе* ***роденовского*** *мыслителя...; ... концертные спуски* ***шопеновских*** *мазурок...; ... гремящий тарелками* ***барбизонский*** *полдень...; ... знаменитых* ***Тацитовых*** *Анналов..; ... аллее, параллельной Малой* ***Бронной*** *улице...; англосаксонский* ***тенишевский*** *стиль...;* ***... Каменноостровский*** *– это легкомысленный красавец...* и т.д.

**НСМ** хорошо справляется и с такими словообразовательными моделями, как превосходная степень прилагательных с суффиксом ***–ейш-*** **:** *... поражен ее невниманием ко мне,* ***чистосердечнейшей*** *естественностью этого невнимания; длиннейшее****,*** ***розовейшее*** *облако,...* (Н)

## Задачи постморфанализа

### 1. Анализ продуктивных словообразовательных моделей

Многие из продуктивных моделей не удается правильно проанализировать только по концам слова. Необходима модель – фрагмент постморфологии, анализирующая такие частые формы, как сравнительная степень с приставкой ***по-*** (***поскорее, поудобнее, побыстрее, поменьше, побольше, подальше***.), и др.: *... в узенькой, стриженой* ***немочке...;*** *, и что* ***близехонько*** *от него* (Н).

### 2. Анализ тривиальных, но недоступных словарному морфанализу случаев,

когда к словам присоединены через дефис частицы *-то* (*я-то, меня-то. дом-то, стоял-то*) или *-ка*, (*дай-ка, построй-ка*).

Отдельный предмет анализа - соединенные дефисом слова типа ***генерад-садовод*** или ***столяр-краснодеревщик.***

### *Морфо-синтаксические проблемы омонимии*

Помимо перечисленного, для успешной работы с линейной структурой текста в «полевых условиях» важно представить спектр возможных омонимий, т.е. на что по параметрам линейной конфигурации похоже явление, чему оно омонимично. Необходимо прогнозировать по линейной структуре потенциальные неоднозначности и знать, в каких ситуациях они возникают и как разрешаются.

### *Проблемы морфологической автономности*

Одна из проблем – выявление словосочетаний, когда важные для синтаксического анализа слова выступают не в обычном своем морфологическом статусе. Это одна из чрезвычайно значимых для анализа текста проблем снятия морфологической неоднозначности при потере словами морфологической автономности.

(1) Такой морфологически очевидный случай как **разрыв неопределенно-личного местоимения предлогом** – ***ни с кем, не для кого*** и т.п. требует лишения словоформ ***кем*** и ***кого*** морфологической автономности – интерпретации этой последовательности словоформ как предложной группы со слугою – неопределенно-личным местоимением (при этом не важно, каким способом это в конкретной системе реализуется). Если мы этого не сделаем, ***кем*** и ***кого*** сохранит значение подчинительного союза – слова, определяющего в процессе сегментации статус сегмента как придаточного предложения.

Однако в предложении ***Я не знаю, ни с кем он ушел, ни для чего взял с собой эти книги...*** группы ***ни с кем*** и ***ни для чего***, внешне тождественные разорванным предлогом местоимениям, на самом деле – именно предложные группы со слугами – подчинительными союзами, а ***ни*** – итеративные операторы сочинения.

Т.о. уже из приведенных примеров видно, что необходим **постморфологический анализ -** морфо-синтаксический анализ разного рода слов и словосочетаний, которые не могут

получить адекватные морфологические характеристики на словарной основе и на основе НСМ.

### 3. Анализ устойчивых словосочетаний

При постморфанализе необходимо строить устойчивые словосочетания, морфологически выступающие как единое целое: 3.1. сложные предлоги; 3.2. группы разорванных вложением предлогов местоимений *никто, ничто, некто, нечто, никакой;* 3.3. объединять группы слов, функционирующие как наречия: они могут иметь в своем составе слова, очень важные при анализе. но в рамках этих устойчивых словосочетаний теряющие свои грамматические свойства; 3.4. имена собственные, 3.5. аббревиатуры, 3.6. названия в кавычках, 3.7. комбинированные формы записи числительных, 3.8. слова и словосочетания, которые могут быть вводными оборотами, и т.д.

Хочется остановиться на нескольких характерных случаях.

***Подчинительный союз \ предикат vs. компонента словосочетания-«наречия»***

Рассмотрим некоторые важные случаи таких неоднозначностей.

1. 1. ***ЛИ:*** слово ***ли*** в сегменте является «сегментообразующим», т.е.показателем того, что содержащий его отрезок, ограниченный запятыми, – минимальная компонента сегмента – придаточного предложения (***Я не знаю, придет ли он. Я не знаю, ему ли ту шапку, что кто-то вчера здесь забыл, отдали.)***.

Но в предложениях ***Мальчик, вряд ли понимающий, чего от него хотят, молчал*** и ***Едва ли он может это сделать*** слово ***ли*** не должно быть проинтерпретировано как сегментообразующий элемент.

2. 2. ***КАК:*** аналогично тому, что говорилось о ***ли***, ***как*** (омоним с довольно сложной функциональной синтаксической парадигмой), теряет сегментообразующую способность, оказываясь компонентой некоторых устойчивых словосочетаний. Например, *Он* ***как раз*** *этим занимается. Он задел его* ***как бы*** *случайно.*

У Мандельштама: *На вопрос Юлий Матвеевич издавал странный грудного тембра неопределенный звук,* ***как бы*** *извлеченный из трубы неумелым музыкантом.* Последний пример показывает, что может произойти, если на уровне постморфанализа не лишить ***как*** статуса подчинительного \ сравнительного или какого-либо другого союза, имеющего функцию сегментообразования.

3. 3. Предикат ***СТАЛО*** в словосочетании ***во что бы то ни стало*** теряет свой морфологический статус предиката: *Он* ***во что бы то ни стало*** *хочет....;* аналогично предикат ***РАВНО*** в словосочетании ***все равно*** в одном из омонимичных значений перестает быть предикатом: *Он* ***все равно*** *этого не выучит * ***Все равно*** *этого не миновать* vs. *Значение коэффициента* ***равно****...\ При любых значениях аргумента все* ***равно*** *пяти.*

*Что* ***угодно*** – словосочетание – омоним: ... [*избегает объяснений, говоря что* ***угодно****..., делая* ***что угодно..., думает при этом о чем угодно..***.] vs.[*делает, что* ***угодно*** *начальству*...]

4. 4. Подчинительный союз ***ЧТО*** в тех же словосочетаниях***,*** а также в словосочетаниях ***только что (только что вошел), пока что (он пока что ничего не сказал)*** и ***хотя*** в словосочетании ***хотя бы (***теряет значение подчинительного союза.

5. 5. 5. ***ЛИ*** в сочетаниях *вряд* ***ли,*** *едва* ***ли*** теряет морфологическую автономность и тем самым - способность вводить придаточное. Аналогично ***ХОТЯ*** в словосочетании ***хотя бы*** теряет значение уступительного союза.

Анализ списка устойчивых словосочетаний, где значимые для анализа слова не могут выступать как подчинительные союзы, предикаты или другие синтаксически важные части речи, необходимо должен быть включен в постморфологический анализ до начала

собственно синтаксического анализа ( к сожалению, такой список при разработке каждой системы составляется заново).

Эти предварительные этапы анализа важны, в частности, и для построения именных и предложных групп, так как в норме предикат и подчинительный союз существенны при поиске хозяина атрибутивного определения.

***Морфологический анализ аббревиатур, имен собственных людей и названий в кавычках***

Самым простым является опознавание аббревиатур (они могут быть и в кавычках). Остановимся на двух других проблемах: морфологической идентификации имен собственных и названий в кавычках.

***Названия в кавычках (ИК)***

Каков морфологический статус слов в названиях «Убить пересмешника», «Любить», т.е. названий книг, спектаклей, фильмов и т.д. – практически любого вида словосочетаний в кавычках, кроме цитаты или прямой речи?

Возникает два знаковых уровня: слова оказываются успешно обработаны модулем автоматического морфанализа, и как единицы словосочетания они соответствуют своему обычному морфологическому статусу, но в тексте элементы этих словосочетаний перестают быть минимальными автономными единицами. Автономное морфологическое единство в предложении эти словосочетания представляют собою только как единое целое. Т.о. приписывание морфологических характеристик таким словосочетаниям – названиям, выделенным кавычками или иным способом – относится к уровню морфологического анализа.

***Имена собственные (ИС)***

Морфанализ ИС – имен людей - составляет существенную проблему, так как цепочки ИС после описанного двухэтапного морфанализа (словарного и НСМ) получают немыслимые морфологические характеристики, создающие колоссальный синтаксический «шум». ИС, включающие иностранные имена и фамилии, требуют или составления особого, постоянно пополняемого, словаря ИС со всеми возможными вариантами, **что практически не реализуемо**, или же – отдельного алгоритма их распознавания. В описываемой системе разработана первая версия такого алгоритма.

Распознавание ИС чрезвычайно сложно. Их анализ проводится, как для любых рекурсивных структур, справа налево: конструкции с ИС, аббревиатурами и ИК часто образуют сложные матрешки: ***обозреватель газеты "Новости недели" Аскар Умаров; спецкор "МН" Валерий Батуев; книгу покойного барина Проферансова "Психический магнит"; ведущий проекта "Формула власти" Михаил Гусман*** и т.д.

Идентифицируя некоторую последовательность потенциальных ИС (с учетом того, что географические названия заданы в словаре), невозможно определить ни падежа, ни рода, ни числа имени собственного из-за немыслимого разнообразия транслитерируемых иностранных имен. Так как при их анализе не удается опираться ни на окончание, ни на «хвосты», их приходится приравнивать к неизменяемым существительным типа ***пальто*** или ***какаду***, с учетом, что они м.б. и мужского, и женского рода: задавать им все падежи мужского и женского рода и множественного числа.

Легко распознаются только конструкции типа М.И.Павлов, Дж.Буш, Ф.Тефт, Р.Паксасу, А.Паулаускас, т.е. конструкции, начинающиеся с заглавных букв с точкой.

Сложности возникают, например, при анализе цепочки, примыкающей к первому слову предложения. Даже если оно опознано при словарном морфанализе, мы вынуждены подозревать в нем ИС (***Тони Блэр,*** например)

Мы почти никогда не можем быть уверены в том, как именно несколько идущих подряд слов с заглавных букв (а эта задача возникает, начиная с цепочки из двух компонент) следует объединять в цепочки - группы, соответствующие именам разных лиц. Эти ситуации в пределах одного предложения порождают неоднозначности, разрешимые лишь при расширении контекста.

В разрабатываемой системе предусмотрены особые правила, формирующие конструкции с именами собственными, в которых учитываются возможные неоднозначности, обусловленные тем, что их падеж непредсказуем, и, в частности, правила, определяющие поверхностно-синтаксического «хозяина» имени собственного, что необходимо для сегментации.

Комбинаторные сложности возникают, например, в случаях, когда существительное – потенциальный хозяин ИС - относится к классу слов, имеющих объектную валентность (как в конструкциях типа **субъект**= ***учитель, наставник, дед, брат***..+. **объект** и далее следует цепочка ИС (например, ***X Y)***.

Даже для русских ИС такие конструкции потенциально омонимичны: конструкция ***учительница Евгения Шульгина*** может быть понята и как (1) учительница (чья?) – человека, которого зовут Евгений, фамилия ее – Шульгина, и как (2) учительница, которую зовут Евгения Шульгина, так как мы не знаем заранее, (А) заполнена ли объектная валентность (она не сильная) и (Б) если она заполнена, то где проходит граница между именем объекта и именем субъекта.

Трудности порождают и другие специфические конструкции, в рамках узкого лингвистического контекста не дешифруемые, например, ***исполняющий функции директора Исакадзе*** и т.д.

Уже упоминались сложности ситуаций, когда ИС совпадает с каким-то обычным словом и к тому же является первым словом в предложении. Сколько значений, например, должно быть приписано словосочетанию ***Рой Медведев..***. (***рой*** – омоним по результатам словарного морфанализа: (!) существительное и (2) повелит. накл. от ***рыть,*** а заглавная буква в начале предложения дает нам возможность понять ***Рой*** и как имя собственное, т.е. возникают три способа понимания.

## Заключение

В пределах данного объема текста нет возможности подробно разобрать все аспекты реальных сложностей морфологического и морфосинтаксического анализа. Помимо перечисленных, существует задача формирования числительных в комбинированной буквенно-цифровой форме (***двести шестой; один млн 236 тыс ; 16 сотен 25 десятков и т.д.***) Хочется отметить, что для любой системы, ориентированной на ПСА «живого» текста, все названные задачи неизбежно приходится решать, и было бы полезно как можно детальнее исследовать контекстно-комбинаторные варианты для каждого из поставленных вопросов.

## Литература

1. Н. А. Пащенко. Об одном подходе к проблеме снятия омонимии при автоматической обработке текстов на естественном языке. НТИ №4 1967.

2. А. А. Зализняк. Грамматический словарь русского языка. М: Русский язык, 1980.

3. Ж. Г. Аношкина. Словарь омонимичных словоформ русского языка. М: Машинный фонд русского языка Института русского языка РАН, 2001. (http://irlras-cfrl.rema.ru:8100/homoforms/index.htm)

4. Т. Ю. Кобзарева, Д.Г. Лахути, И.М. Ножов. Модель сегментации русского предложения. Диалог'2001. Аксаково 2001. т.2 с.185-194.

5. Т. Ю. Кобзарева, Р.Н. Афанасьев. Универсальный модуль предсинтаксического анализа омонимии частей речи в русском языке на основе словаря диагностических ситуаций. Диалог'2002 Протвино 2002, т.2, с.с.258-268.