О СРАВНИМОСТИ И ЭКВИВАЛЕНТНОСТИ КОМПЬЮТЕРНЫХ ПРЕДСТАВЛЕНИЙ МОРФОЛОГИИ

С.А.Коваль,
(Филологический факультет СПбГУ)

Модули морфологического анализа или (несколько реже) синтеза являются обязательными компонентами всех компьютерных систем переработки текстовой информации на языках с нетривиальной морфологией, к каким, несомненно, относится и русский язык. Разработчики таких систем встают перед необходимостью заниматься созданием морфологического модуля, даже если он выполняет в их приложении далеко не центральную функцию. Несмотря на наличие отдельных традиций и эволюционных линий в практике компьютерной морфологии русского языка, постоянно создаются новые морфологические компоненты, которые функционально повторяют уже имеющиеся. Причиной тому являются трудности в освоении имеющихся готовых разработок, связанные с отсутствием единых принципов составления документации, а также с отсутствием методик определения пригодности имеющихся компьютерных моделей морфологии для новых приложений.

Исходя из того, что оценка морфологического компонента компьютерной системы может быть сведена к сравнению нескольких кандидатов при выборе прототипа или же к сравнению существующего компонента с виртуальным эталоном, определяемым внешними функциональными характеристиками, мы считаем важной задачей формализацию сравнения различных моделей морфологии. В качестве начальных шагов к решению этой задачи предлагается определить точный смысл понятий «представление морфологии», «эквивалентность представлений морфологии», «сравнимость представлений морфологии», на которые можно было бы опереться в дальнейшем при построении сначала математических, а затем и кибернетических моделей сравнения морфологических компонентов.

Для большей наглядности мы будем опираться в последующем на материал фрагмента русской словоизменительной системы и соответствующие ему фрагменты нескольких существующих моделей морфологии. Рамки избранного языкового материала будут определены в третьем разделе.

С точки зрения выполняемых практических функций мы ограничимся моделями, ориентированными на обработку письменного текста на русском языке.

1. Лексикон и представление как компоненты модели морфологии

При имеющемся разнообразии алгоритмов анализа и синтеза более перспективным выглядит повторное использование не процедурных (динамических), а статических морфологических ресурсов. Речь идет в первую очередь о словарных базах данных, приписывающих лексическим единицам морфологические (обычно словоизменительные) характеристики. В скрытой форме эти словарные базы данных несут в себе информацию о динамических процессах, с помощью которых можно переходить от одной формы слова к другой. Но, сравнивая реально существующие словарные базы данных, мы будем приводить их к единообразному виду, вынося всю явную процедурную информацию за их пределы. Таким образом, под ***моделью морфологии*** мы будем понимать совокупность ***лексикона***, являющегося статическим массивом данных, где лексемам приписаны их морфологические типы, и ***представления морфологии***, где морфологические типы получают свою расшифровку.

В основу построения лексикона будет положен ***принцип генерализации***: информация о морфологических свойствах некоторой лексемы описывается для нее не индивидуально, а с отсылкой к соответствующему разделу описания (представления морфологии), посвященному такому набору свойств, которые данная лексема в большинстве случаев делит с другими лексемами.

Если в реально существующих словарных базах принцип генерализации оказывается не соблюден, то для сравнения с другими базами их необходимо преобразовать, последовательно проводя этот принцип. Эта процедура (генерализация) не является единственной в процессе ***нормализации*** словарных баз данных, то есть в приведении их к нормальному виду, облегчающему сопоставление и оценку. Другие принципы нормализации будут описаны в следующем разделе.

Что же касается представления морфологии, то эта составляющая модели будет интересовать нас здесь прежде всего как классификация лексем, отражающая их объективные свойства. При этом способ моделирования морфологических процессов в той или иной модели оказывается просто операциональным критерием, применяемым для классифицирования. Это позволяет сравнивать представления морфологии в аспекте, относительно свободном от рассмотрения конкретных реализаций морфологических процессов.

2. Нормализация лексиконов

Конечной целью нормализации лексиконов является облегчение формализации сравнения соответствующих им представлений морфологии. Проводя ряд преобразований, направленных на приведение лексиконов к единообразному виду[1], мы тем самым эксплицируем и классификацию лексики по морфологическим характеристикам, используемую в данной модели.

Содержательно нормализация лексикона должна обеспечить соблюдение следующих принципов:

1) ***принцип достаточности***: информации, поставленной в соответствие лексеме в для описания ее морфологических характеристик, должно быть достаточно для установления всех ее словоформ;

2) ***принцип неизбыточности***: к описанию морфологического типа относится только та информация, которая используется для порождения хотя бы одной из словоформ лексемы; разумеется, в реальных словарных базах лексемам может приписываться разнообразная информация, не подпадающая под этот принцип, однако эта информация не должна рассматриваться как часть модели морфологии;

3) ***принцип однозначности*** в приписывании морфологического типа; в терминах теории баз данных речь идет об устранении множественных полей, что осуществляется путем разделения записей с множественными полями на несколько записей; это может приводить к появлению записей с совпадающими (квази)основами, однако в нормализованной базе должен использоваться составной ключ, состоящий из (квази)основы и морфологического типа;

4) ***принцип генерализации***, который был введен в предыдущем разделе; проведение этого принципа заключается в атомизации и инвентаризации всех приписанных морфологических типов, то есть все неэлементарные описания морфологических характеристик должны быть переобозначены как элементарные, а затем сведены в единый список с неповторяющимися элементами, который станет основой представления морфологии для данной модели.

3. Иллюстративный материал

3.1 Фрагмент русской словоизменительной системы

Уже введенные и вводимые далее определения будут иллюстрироваться на фрагменте системы русского именного словоизменения, ограниченной следующими различительными признаками по [1, с. 291]:

- грамматический разряд - субстантивный;
- морфологический класс - 1;
- тип склонения - I субстантивный;
- нестандартная омонимия чисел;
- морфонологический тип условной основы 1.

В системе обозначений, принятой в словаре [2], это соответствует приписыванию слову индекса, начинающегося с фрагмента «м 1», то есть речь идет о неодушевленных существительных мужского рода, относящихся к школьному

[1] По сути речь идет здесь об устранении из лексикона множественных и неключевых полей, так что результирующая таблица будет соответствовать требованиям ко всем четырем нормальным формам реляционных баз данных.

второму (вузовскому первому) склонению, не оканчивающихся на мягкий согласный, шипящий, заднеязычный или *ц* (например, *комбайн*, *мир*, *дуб*, *лес*, *глаз*). Всего в электронной версии словаря А. А. Зализняка, представленной на сайте С. А. Старостина (http://starling.rinet.ru/downlru.htm), насчитывается почти 8000 таких существительных.

3.2. Рассматриваемые модели морфологии

Понятия, вводимые здесь для сравнения моделей морфологии, могут быть проиллюстрированы с обращением к ряду современных моделей русской словоизменительной системы. В частности, были рассмотрены описания представлений морфологии, загруженные с сайта А. В. Поминова «Мультитран» в январе 2003 г. (http://www.multitran.ru/RussianMorphologyClasses.htm; далее - «модель "Мультитран"») и с сайта «Автоматическая обработка текста» инициативной группы разработчиков программного обеспечения с лингвистическими компонентами в ноябре 2002 г. (http://www.aot.ru/download.htm; далее - «модель АОТ»). Анализировалось также описание структуры электронной версии словаря А. А. Зализняка, созданной в Украинском языково-информационном фонде Национальной академии наук Украины, которое было любезно предоставлено нам Т. А. Грязнухиной и Т. П. Любченко в ноябре 2002 г. (далее - «модель Фонда»)

Кроме того, привлекались данные описания морфологического анализатора О. С. Кулагиной в [3] и описание стандартных морфологических объектов и фрагменты словаря системы ЭТАП по состоянию на март 2002 года, за предоставление которых мы весьма признательны Л. Л. Иомдину.

Во всех названных моделях морфологии построение словоформ на базе основы моделируется только с помощью механизма конкатенации этой основы со словоизменительными формативами. При этом в трех первых моделях каждой лексеме отводится, как правило, одна запись в словарной базе данных, поэтому при наличии чередований вместо обычной для фундаментальной морфологии основы в словарную базу вписывается ее начальный отрезок, свободный от чередований, - квазиоснова (*уз* для *узел*, *псал* для *псалом*, *р* для *ров*). Это приводит к увеличению числа морфологических типов (за счет удлинения квазиокончаний), но позволяет упростить процедуру морфологического анализа. В двух других моделях чередования в основе моделируются с помощью процедуры выбора чередующегося варианта (организованной в двух моделях несколько различающимися способами), что позволяет оперировать такими основами, которые максимально приближены к словоизменительным основам фундаментальной морфологии.

Отличительной особенностью модели «Мультитран» является учет буквы *ё*, то есть здесь русский язык описывается в несколько иной форме, чем в других моделях. В целях более контрастного проявления других отличий данной модели можно заменить букву *ё* в списках квазиокончаний этой модели буквой *е*, что позволит исключить некоторые морфологические типы. Другой особенностью является наличие двух полей для В. падежа как в ед., так и во мн. числе, что подразумевает потенциальную возможность каждого существительного использоваться и как одушевленное, и как неодушевленное. Для рассматриваемого фрагмента лексики эти два дополнительных поля можно считать избыточными. С их исключением число различающихся морфологических классов станет еще меньше.

В модели АОТ для географических названий, наименований организаций и личных имен людей вводятся особые парадигмы, которые могут совпадать с уже имеющимися парадигмами для имен нарицательных. Признак принадлежности существительных к одному из названных разрядов не играет никакой роли при образовании словоформ данной лексемы и должен быть устранен при проведении принципа неизбыточности представления.

Представление рассматриваемого фрагмента русской морфологии в модели Фонда содержит четыре пары парадигм, внутри которых различия наблюдаются только в порядке следования вариантов квазиокончаний. Такая организация представления обеспечивает, наряду с распознаванием нескольких возможных вариантов отдельных словоформ, выбор более распространенного варианта при синтезе[2] (с приведенной в разделе 4 оговоркой относительно дистрибуции основного и второго П. падежей). Тем не менее, отметив это важную особенность модели Фонда, мы в дальнейшем не будем ее учитывать, чтобы нагляднее представить другие отличия между моделями морфологии.

Сложнее провести нормализацию словарной базы данных в модели Кулагиной. Принцип однозначности требует расщепить словарные статьи, отведенные основам и/или словоформам сразу нескольких лексем (например, статья с ключом *мест* несет в себе информацию о лексемах *место* и *месть*) Принципы достаточности и генерализации требуют рассматривать в качестве морфологических типов всевозможные комбинации того, что у О. С. Кулагиной названо словоизменительным типом, и того, что названо категорией заголовка. Первое понятие характеризует набор окончаний полной парадигмы. Всего в Кулагина 1986 выделено для рассматриваемого фрагмента лексико-морфологической системы 11 словоизменительных типов. Второе понятие (чрезвычайно близкое понятию «маски»

[2] См. об этом также Грязнухина и др. [4, с. 66].

в морфологическом компоненте системы ЭТАП) задает распределение дефектных полей в парадигме. Поскольку в работе О. С. Кулагиной не приводится ни список этих комбинаций, ни полный список «категорий», дальнейшее сопоставление этой модели с другими в рамках настоящей работы невозможно.

Нормализация морфологической модели системы ЭТАП - еще более сложная задача, подробное рассмотрение которой значительно увеличило бы объем настоящего доклада. Укажем лишь два главных осложняющих фактора. Во-первых, словарная база данных присутствует в этой модели, если опираться на описание в [5, с. 53-55], в двух вариантах: до процесса, называемого «трансляцией», когда база индексирована по идентификаторам лексем, и после трансляции, когда в качестве ключа поиска по базе выступает основа. Каждая из этих ипостасей задает свою классификацию морфологических типов, свое представление морфологии (которые, разумеется, эквивалентны друг другу в смысле, определяемом далее). Во-вторых, описание морфологических характеристик здесь не генерализовано, а записывается в виде выражений на достаточно сложном метаязыке, которому свойственны такие черты естественного языка, как синонимия и свободный порядок слов. Поэтому построение представления морфологии по предложенным нами принципам возможно только после разработки процедуры нормализации выражений, описывающих морфологическую информацию, при наличии доступа к словарной базе данных в полном объеме.

4. Сравнимость представлений морфологии

Чтобы можно было ставить вопрос о сравнении двух представлений (и всех моделей, в которые они включены), нужно, чтобы они выражали грамматическую информацию о словоформах на едином метаязыке, по крайней мере, на метаязыках, легко сводимых друг к другу за счет простых переобозначений. При этом, например, результат сравнения двух представлений может не зависеть от того, одинаково ли трактуется в них частеречная принадлежность причастий, если ввести соответствие типа: «морф. класс = глагол; репрезентация = причастие» в одном представлении типу «морф. класс = причастие» в другом представлении.

Иначе говоря, два представления (модели) морфологии ***сравнимы***, если существует взаимно однозначное соответствие между множествами комбинаций граммем, которые могут быть приписаны в сопоставляемых представлениях отдельным словоформам.

Обратимся к нашему материалу.

Большинство анализируемых представлений позволяют описывать словоформы в рамках рассматриваемого фрагмента лексикона с помощью одной из 12 возможных комбинаций граммем. Если отвлечься от общего для всех этих словоформ маркера части речи - имени существительного, эти комбинации представляют собой пары, состоящие из граммемы числа (два различных значения) и из граммемы падежа (шесть различных значений). При этом формы второго Р. (*сахару* в *ложка сахару*) и второго П. (*пруду* в *на пруду*) падежей рассматриваются как вариантные для, соответственно, Р. и П. падежей. Подобное решение вполне приемлемо для компьютерных систем, ориентированных на анализ текста, поскольку обеспечивает адекватную интерпретацию входных словоформ. В части трактовки второго Р. падежа такое решение можно признать приемлемым и для систем, обеспечивающих синтез русского текста при том условии, что порядок задания вариантов является существенным, и при наличии более одной возможности словоформа синтезируется согласно правилу, указанному первым. Так, в качестве формы Р. падежа от существительного *сахар* всегда будет строиться форма *сахара*, как в контексте *у сахара*, так и в контексте *ложка сахара*, если правило для образования этой формы будет предшествовать правилу для образования *сахару*.

В отличие от второго Р. падежа второй предложный не составляет с какой-либо иной формой набор варьирующихся вариантов. Основной П. падеж, который отождествляется с ним в большинстве рассматриваемых представлений, находится с ним в отношении дополнительной дистрибуции. Поэтому неразличение первого П. и второго П. падежей не может адекватно обеспечить данными процедуру синтеза. В лучшем случае алгоритм синтеза, строящийся на этом представлении, будет включать ad hoc правило со смыслом «если данное существительное может образовывать форму П. падежа ед. числа неединственным образом, после предлогов *в*, *на* выбери форму, оканчивающуюся на *у*, *ю*». В худшем случае синтез будет строить одни и те же формы как для первого, так и для второго П. падежа недифференцированно, создавая ошибочные формы типа *о пруду* либо *в пруде*.

Из всех упомянутых моделей морфологии только модель системы ЭТАП рассматривает второй Р. и второй П. падежи как отдельные граммемы. Таким образом, количество возможных комбинаций граммем для нашего фрагмента лексикона в этой модели будет превышать 12, и представление, которое можно построить для этой модели, окажется несравнимым с остальными. Это означает также, что представление рассматриваемого фрагмента русской словоизменительной морфологии в модели ЭТАП лучше подготовлено для использования в морфологическом синтезе.

5. Эквивалентность представлений морфологии

Прежде чем определить понятие эквивалентности для представлений морфологии, введем понятие мощности модели морфологии.

Мощность модели морфологии характеризуется множеством словоформ, которые в ней могут быть построены, и приписанными этим словоформам наборами граммем. Разумеется, сравнение двух моделей по мощности может быть трудно реализуемым на практике, однако оно не является невыполнимым для конечных лексиконов и достаточно формализованных представлений морфологии.

Естественно, проводя сопоставление двух моделей морфологии по их мощности, следует провести все необходимые переобозначения граммем, которые упоминались выше в связи с определением сравнимости. Если окажется, что множества порождаемых словоформ в двух моделях тождественны и после переобозначения одним и тем же словоформам соответствуют одни и те же наборы граммем, можно сказать, что эти две модели морфологии имеют одинаковую мощность.

Разумеется, в реальности такая ситуация едва ли когда-либо встретится, поскольку вероятность того, что две независимо создаваемые модели морфологии покрывают один и тот же набор лексем, ничтожно мала. Однако при определении эквивалентности двух представлений нас будет интересовать не фактическое равенство их по мощности, а то, какими средствами можно этого равенства достичь. Отметим, что, установив неравномощность двух моделей, далеко не всегда можно определить, у какой из них мощность больше: превосходя другую модель в отражении одних морфологических типов, эта же модель может уступать ей в отражении других типов[3].

Будем говорить, что два представления (и все построенные на них модели) ***эквивалентны***, если на их основе можно построить модели одинаковой мощности только за счет наращивания соответствующего лексикона. Иначе говоря, эквивалентность двух представлений понимается как их потенциальная способность охватить один и тот же круг морфологических явлений независимо от того, с какими лексиконами работают эти представления в исходном состоянии модели. Чтобы продемонстрировать эту эквивалентность в явном виде, необходимо дополнить лексикон первой модели теми лексемами, которые изначально были только в лексиконе второй модели, и наоборот, после чего можно будет убедиться в одинаковой мощности результирующих моделей.

На практике нет необходимости проводить эту трудоемкую и ресурсоемкую для вычислительной техники процедуру от начала до конца, достаточно проанализировать наборы морфологических типов в обоих представлениях и установить между ними соответствия. При этом эквивалентность, вообще говоря, не имеет необходимым условием равное количество морфологических типов.

Рассмотрим с этой точки зрения три представления морфологии, соответствующие выделенным фрагментам нормализованных лексиконов в модели «Мультитран», модели АОТ и модели Фонда. Ограничения на объем настоящей публикации не позволяют привести все выявленные соответствия и несоответствия, поэтому мы ограничимся перечислением наиболее существенных отличительных особенностей этих трех моделей.

Только модель «Мультитран»:

- не позволяет запретить генерацию форм мн. числа (для имен singularia tantum);
- не позволяет задать в явном виде вариантность некоторых словоформ (*лоскутья*/*лоскуты*, *секторы*/*сектора*), хотя не исключает задание этих лексем двумя записями в словарной базе данных.

Только модель АОТ:

- позволяет запретить генерацию форм мн. числа в некоторых морфологических типах с беглым гласным (например, с чередованием *ет*/*т* типа *Египет*);
- позволяет задать варианты окончания Р. падежа мн. числа Ø/*-ов*.

Модель Фонда отличается избирательным отражением некоторых редких словоизменительных типов. Например, только эта модель

[3] Возвращаясь к другим упомянутым ранее моделям морфологии, отметим одно неоспоримое достоинство модели ЭТАП. Гибкость используемого здесь метаязыка описания морфологических типов позволяет дополнять морфологический словарь этой системы до сравнимой модели морфологии как угодно большой мощности. Иначе говоря, отсутствие генерализации оборачивается здесь тем, что морфологическая модель ЭТАП имеет максимально возможную мощность среди всех сравнимых с ней моделей. Оборотной стороной этого достоинства видится закрытость данной модели, то есть затрудненность обмена данными между морфологическим словарем ЭТАПа и словарными базами, построенными на других принципах.

- позволяет породить особую форму второго предложного падежа *во льну* для существительного *лен*;
- не позволяет построить парадигму существительного *кочан* с беглым *а*.

Конечно, для большинства моделей морфологии (в частности, для всех рассмотренных) отсутствие того или иного морфологического типа не является непреодолимым препятствием. Однако предпринятое нами сравнение имело целью показать, какова может быть последовательность действий при оценке моделей в их текущем состоянии с точки зрения возможности переноса из них данных в другие модели. Именно результаты подобного сравнения могут быть использованы для дополнения представления недостающими морфологическими типами и достижения эквивалентности с другими представлениями. Кроме того, во всех анализируемых моделях были выявлены единичные погрешности в представлении языковых фактов (связанные, например, с опечатками при заполнении базы данных), о которых было сообщено разработчикам.

По итогам сравнения можно заключить следующее. Среди трех моделей, подвергнутых нормализации, никакие две не являются эквивалентными друг другу. Разница в мощности этих моделей может быть соотнесена с задачами, для которых они создавались. Так, с учетом всех особенностей моделей, упомянутых в предыдущих разделах, можно утверждать, что к задачам морфологического синтеза лучше всего приспособлена модель Фонда. Модель «Мультитран» представляет собой достаточно мощную модель для морфологического анализа или, скорее, лемматизации (установления лексемы по словоформе без указания грамматического значения последней), например, на входе в компьютерный словарь. Модель АОТ отличается более тщательной проработкой некоторых периферийных явлений словоизменительной системы, что позволяет брать ее за основу при построении анализаторов, работающих с разнородными входными текстами.

Таким образом, результаты сравнения моделей морфологии могут использоваться для установления того, в какой мере особенности представлений обусловлены функциональным назначением.

Заключение

Выше были предложены определения ряда понятий, которые могут быть положены в основу при разработке формальных методик сравнения моделей морфологии и установления их пригодности для повторного применения. Разумеется, введенные здесь понятия модели морфологии и ее составляющих, сравнимости и мощности моделей, эквивалентности представлений морфологии нуждаются в дальнейшем уточнении, однако уже сейчас их можно было применить для сравнения нескольких моделей морфологии. Это помогло выявить существенные особенности каждой модели, обусловленные, как правило, их специфическим функциональным предназначением, а также обнаружить отдельные погрешности, о которых было сообщено разработчикам.

Можно предположить, что сравнение двух представлений может происходить со значительно меньшими затратами труда и времени, если за точку отсчета взять всеобъемлющее генеральное представление русской словоизменительной морфологии, которое еще предстоит создать[4]. В этом случае сравниваемые представления могут быть рассмотрены как подмножества генерального представления, в котором нейтрализованы некоторые различительные признаки морфологических типов.

Литература

1. Зализняк А. А. Русское именное словоизменение. М.: Наука, 1967.
2. Зализняк А. А. Грамматический словарь русского языка. Словоизменение. М.: Русский язык, 1977. (2-е изд., испр.: 1980; 3-е изд., испр.: 1987; 4-е изд., испр. и доп.: 2003).
3. Кулагина О. С. Морфологический анализ русских именных словоформ // Препринт № 10 / Институт прикладной математики АН СССР. М., 1986.
4. Грязнухина Т. А., Любченко Т. П., Рабулец А. Г. Электронная версия Грамматического словаря русского языка А. А. Зализняка как инструмент автоматического морфологического анализа русского текста // Корпусная лингвистика и лингвистические базы данных. Доклады научной конференции. 5-7 марта 2002 г. Санкт-Петербург. СПб: Изд-во С.-Петерб. ун-та, 2002. С. 63-70.
5. Апресян Ю. Д. и др. Лингвистический процессор для сложных информационных систем. М.: Наука, 1992.

[4] В этом направлении развивается проект, описанный в [6].

6. Коваль С. А. К унификации представления русской морфологии в системах обработки текстовой информации // Компьютерная лингвистика и интеллектуальные технологии. Труды Международного семинара Диалог'2002. Т. 2. М.: Наука, 2002. С. 269-275.