А.В.Венцов, Е.В. Грудева, В.Б.Касевич, Е.В.Ягунова

Идиомы в Национальном корпусе русского литературного языка

Необходимость учета идиом возникает при решении задачи морфологической разметки текстов в рамках создания репрезентативного корпуса языка. Одновременно эта проблема оказывается существенной и с точки зрения моделирования восприятия речи, поскольку восприятие естественным образом опирается на словарь и необходимо определить, что является единицей словаря. Как можно видеть, теоретические и прикладные аспекты здесь лишь с трудом поддаются разграничению.

Поставленная проблема – проблема выделения идиом, их обработки в различных прикладных целях – не имеет единственного, однозначного решения. Мы рассмотрим разные подходы к решению соответствующих проблем, намечая возможные варианты ответов на вопросы (ср. Мустайоки, Копотев 2004).

Обращает на себя внимание наличие существенной рассогласованности в описании понятия идиомы в стандартных источниках устоявшегося лингвистического знания – в словарях и энциклопедиях. Например, в словаре лингвистических терминов (Ахманова 1966) понятие идиомы трактуется со стилистических позиций («фразеологическая единица, обладающая ярко выраженными стилистическими особенностями, благодаря которым ее употребление вносит в речь элемент игры, шутки, нарочитости» (Ахманова 1966: 165-166)). Приводимое в качестве примера сочетание *приказать долго жить* уже с трудом подходит под определение – не говоря об отсутствии в определении собственно лингвистической трактовки.

В «Лингвистическом энциклопедическом словаре» (Ярцева 1990) «идиома» не представлена в виде отдельной статьи и получает толкование в составе статьи «фразеологизм» (единицы, «характеризующиеся переосмыслением их лексико-грамматического состава и обладающие целостной номинативной функцией («выносить сор из избы», «по горячим следам», «белая ворона»)» (Телия 1990: 559)). Если «переосмыслить» приведенное толкование, то мы придем к тому, вероятно, что, например, в сочетании *белая ворона* его референт не является вороной и не характеризуется как белый, т.е. значение словосочетания *белая ворона* нельзя представить как сочетание значений его слов-компонентов (в отличие, например, от словосочетания *белый гусь*, где мы непосредственно выводим значение сочетания из значений его компонентов). Невыводимость значения сложного, составного целого из значений компонентов (или, иначе, **неаддитивность** семантики идиоматического словосочетания) и есть, насколько можно судить, наиболее широко распространенное понимание идиомы, которое почему-то не проникает в справочную литературу.

Можно предложить и более формальное определение идиомы. Критериями идиоматичности будем считать следующие признаки. Идиоматичность фиксируется там, где а) по крайней мере одно из слов не употребляется вне данного сочетания (*бить баклуши*); б) в рамках идиоматического сочетания нарушаются правила управления или согласования (*в течение*; об этом примере см., впрочем, ниже); в) ни одна из словоформ в составе идиоматического сочетания не может быть опущена без нарушения его семантики – возможно, за вычетом семантики опущенного слова – и функций (ср. *она все равно не узнает* ⇒ **она все не узнает* ⇒ **она равно не узнает*) иначе говоря, *все равно* является идиомой, в отличие от *поодаль от* (ср. *она села поодаль от Петра* ⇒ *она села поодаль*, т.е. *поодаль от* не является идиомой).

Уже примеры, использованные выше, показывают существенную неоднородность словосочетаний, которые по крайней мере некоторые исследователи относят к идиомам, и соответственно возможность существенно разных подходов к решению этой проблемы. Мы остановимся лишь на двух аспектах соответствующей проблематики. **Первый** – это возможность идиоматической и неидиоматической трактовки для одного и того же сочетания. **Второй** – соотношение категорий слова, фонетического слова и идиоматического словосочетания.

Первый аспект

Для рассмотрения неоднозначности трактовки подошло бы и рассматриваемое выше сочетание *белая ворона*, но для большей демонстративности возьмем близкий пример: *белый медведь*.

Отличие заключается в том, что в этом последнем случае нет метафоричности. Сходство же – в неаддитивности семантики: *белый медведь* – это не медведь белого цвета, это особый вид медведя, оппозиция здесь не *белый медведь* vs. *черный медведь* vs. *красный медведь* и т.д., а *белый медведь* vs. *бурый медведь* vs. *гризли* и т.д. Но сочетание *белый медведь* может использоваться и в ситуации, когда речь идет, например, об изображении медведя, выполненном из материала белого цвета (ср. *Вот этого белого медведя, из уральского камня, поставь поближе, а черного, из обсидиана, подальше*).

Еще более ясно неоднозначность сочетания видна в таких примерах, как *Большой театр*: вне смыслового контекста невозможно определить, имеем мы дело с именем собственным или же нарицательным, где речь идет о физическом размере здания театра, его труппе и т.п. Поэтому при решении задач моделирования восприятия речи («поверхностного восприятия», где учет контекста минимален), сегмент (звучащего) текста *большой театр* либо не должен квалифицироваться как идиома, либо, при наличии в словаре отдельной статьи (единицы) *Большой театр*, должны быть предусмотрены два варианта интерпретации: *Большой театр* и *большой театр* (аналогично для *белый медведь и белая ворона*).

Еще один пример – сочетание *друг друга*. Это традиционная идиома, лишь как целое выполняющее функции особого реципрокального

местоимения[1] (anaphore в терминологии Хомского). Словарь должен содержать отдельную статью (единицу) *друг друга*. Но и здесь налицо двузначность, ср. *Там друг друга всегда поддерживают* и *Там друг друга всегда поддержит*.

Но даже и там, где перед нами – «стопроцентная» идиома, не имеющая того, что можно было бы назвать неидиоматическим омонимом (типа *друг друга*), имеет смысл допускать возможность двоякой интерпретации. Например, уже приводившееся сочетание *приказал долго жить* вне всякого сомнения является идиомой и должно войти в словарь на правах отдельной статьи (единицы). Однако едва ли это сочетание, при всей его безусловной идиоматичности, является неанализируемым в абсолютном смысле.

Дело в том, что, по данным некоторых исследований, в ментальном лексиконе человека имеются связи между определенными морфемами или блоками морфем – например, одинаковыми корнями или аффиксами в составе **разных** слов, что создает некую кумулятивную частотность соответствующих морфем. Тем более уместно признать наличие аналогичных связей между словами, входящими в разные сочетания, в том числе идиоматические. Можно предположить, что идиоматичность не блокирует связи тех слов, которые «попадают» в идиоматические сочетания.

Следовательно, при морфологической разметке текста в корпусе русского литературного языка идиомы есть смысл представлять двояким образом (наподобие того, как это сделано финскими коллегами в корпусе ХАНКО (Копотев, Мустайоки 2003)).

Во-первых, должно быть представление идиомы как целого (формально – в качестве сегмента внутри скобок определенного типа, которые и обеспечивают неделимость). При таком представлении идиома выступает как эквивалент слова (целостная единица), и на нее распространяется требование грамматической идентификации, которая в случае идиомы сводится к указанию на частеречную принадлежность (например, *сломя голову* квалифицируется как глагол (или деепричастие) или, возможно, как наречие).

Во-вторых, скобки, выделяющие идиомы, можно и нужно раскрывать, тогда словоформы, входящие в состав идиомы, должны быть снабжены грамматическими характеристиками (как если бы эти словоформы были самостоятельными, не связанными рамками идиоматического сочетания). При этом возникает проблема, каким образом в описании должны отражаться лексемы и словоформы, не представленные вне идиом (ср. *сломя* из *сломя голову* или *баклуши* из *бить баклуши*). Чтобы не нарушать принцип безостаточности индексации единиц текста, мы приходим к необходимости введения фиктивных лексем и/или словоформ (разумеется, с введением таких лексем (словоформ) в словарь).

[1] Здесь можно заметить, что не принятые (к счастью) предложения по введению новой редакции «Свода правил русского правописания. Орфография. Пунктуация» предусматривали для сочетания *друг друга* дефисное написание, т.е. *друг-друга*.

Какие реальности отражает этот возможный подход? Кроме предполагаемых связей совпадающих лексем вне зависимости от их вхождения/невхождения в состав идиоматических сочетаний (см. об этом выше), здесь можно было бы говорить также о специфическом отражении частотности словоформ и лексем. Кроме того, и здесь мы рассматриваем возможные единицы и процедуры поверхностного восприятия.

Идиомы, фразеологизмы нередко выделяются как единицы, эквивалентные слову. Недавно вышедший словарь Р.П.Рогожниковой так и озаглавлен: «Толковый словарь сочетаний, эквивалентных слову: Около 1500 устойчивых сочетаний русского языка»[2] (Рогожникова 2003). Чаще всего при этом имеется в виду, что эквивалентность слову есть семантический признак: устойчивые сочетания, идиомы, фразеологизмы обладают единым неаддитивным значением, уподобляясь в этом отношении слову. Однако единство с семантической точки едва ли может быть операциональным признаком. Трудно определить, какое значение является «цельным», а какое – нет. Как мы видели выше, толкование сочетания *белая ворона* предполагает опору на целый ряд выделимых семантических элементов, и трудно сказать, насколько «спаяны» эти элементы в некое постулируемое целое и присутствует ли такая семантическая целостность вообще.

В предисловии к словарю верно отмечается важность учета формальных признаков «устойчивых сочетаний», которые придают **сочетанию** цельность и тем самым сближают его со словом (Рогожникова 2003: 4-5). По-видимому, этот подход стоит использовать более последовательно. При решении вопроса «слово или словосочетание?» имеет смысл обращаться именно к формально-грамматическим признакам, критериям, которые отличают слово от словосочетания и которые выступают как универсальные, а не изобретаются применительно к каждому отдельному типу сочетаний (Касевич 1977). Один из наиболее важных и употребимых критериев для решения вопроса о том, имеем ли мы дело со словом или словосочетанием, это **критерий вставимости**, где под вставимостью имеется в виду возможность/невозможность вставки так называемых фразовых отделителей (слов, словосочетаний, которые способны выступать как высказывания-фразы) между компонентами сочетания.

Например, *два метра* является словосочетанием, поскольку возможна вставка: *два метра* ⇒ *два погонных метра* (при этом *_погонных* может выступать как реплика-фраза, ср. *Каких метров? – Погонных*).

В качестве другого важного признака, существенного преимущественно для флективных языков, широко используется **раздельнооформленность** (*два метра* и по этому признаку оказывается

[2] Приводимое в подзаголовке число устойчивых сочетаний, скорее всего, изменится, если мы систематически проверим все сочетания с точки зрения их соответствия тем критериям, о которых говорится в данной статье. Но даже если рассматривать эту количественную оценку как приблизительную и ориентировочную, она указывает на высокую степень значимости проблемы идиом: само по себе число идиом сопоставимо, например, с числом корней в словах русского языка.

словосочетанием, а не словом, ср. *двух метров, двумя метрами* и т.п., где каждый из компонентов сочетания получает собственное (морфологическое) оформление).

Обратимся с этой точки зрения к таким часто фигурирующим в литературе примерам идиом (устойчивых сочетаний, эквивалентных слову), как *без умолку, в обнимку, в течение*. Во всех сочетаниях этого типа невозможна ни вставка фразового отделителя, ни демонстрация раздельнооформленности; отрицательный ответ дает и использование других критериев, которые здесь не обсуждались (возможность/невозможность изменения порядка компонентов, наличие/отсутствие самостоятельных синтаксических связей, актуальных или потенциальных). Следовательно, указанные сочетания, скорее всего, являются **словами** с формально-грамматической точки зрения. Их составной характер – признак орфографический, а также этимологический. Но тогда мы должны признать, что это **не** идиоматические сочетания, поскольку они **не неоднословные** единицы.

Иначе говоря, среди единиц, традиционно трактуемых как идиомы, есть слова – которые в силу этого не должны включаться в класс идиоматических словосочетаний – и словосочетания-идиомы типа *белая ворона* и т.п. (где налицо раздельнооформленность).

Какие выводы можно сделать из вышеизложенного? Естественно, признак наличия/отсутствия вставимости и раздельнооформленности (разрывная или неразрывная единица) является важным операциональным и классификационным признаком и должен отражаться в требовании соответствующей грамматической идентификации идиоматической единицы в разметке Корпуса.

Главный вывод: **идиомами** (в широком понимании традиционного подхода) называются принципиально разные единицы, имеющие отражающую это различие грамматическую идентификацию: во-первых, **идиоматические словосочетания** (неоднословные целостности), не выводимые по правилам, а потому включаемые в словарь на правах отдельной единицы; во-вторых, **составные слова** (так называемые «сочетания, эквивалентные слову»). Последними имеются в виду единицы, которые (типа *в обни+мку, в голова+х, а_то+* и пр.).

Более того, в соответствии с рассматриваемыми критериями среди «составных слов» выделяются разрывные и неразрывные (составные) слова (например, *дру+г__о__дру+ге* – с одной стороны, и *в_обни+мку, на_дыбы+, изо_дня+_ в_де+нь* – с другой).

Введение категории (индексации) «составные слова» позволяет решать вопрос о снятии омонимии на уровне «свободное словосочетание» / «целостная единица» (например, *Дру+г о дру+ге всегда+ позабо+тится* и *Они+ ненави+дели дру+г__дру+га*; *Всё_равно+ о+н не придё+т* и *У него+ получи+лось, что всё+ равно+*).

И снова – какие реальности отражает этот возможный подход? Здесь опять же можно было бы говорить о более адекватном отражении

частотности словоформ и лексем[3], и учете предполагаемых связей совпадающих лексем вне зависимости от их вхождения/невхождения в состав идиоматических сочетаний. Кроме того, и здесь мы рассматриваем моделирование восприятия (возможные единицы и процедуры их связей). Например, можно допустить существование нескольких словарей: «в одном, рассчитанном на поверхностное восприятие речи, сочетание *друг друга* в качестве отдельной вокабулы отсутствует (именно такова ситуация в существующих частотных словарях); в другом, призванном обслуживать "полное" восприятие, включены в качестве отдельных вокабул и словоформы *друг, друга*, и сочетание *друг друга*» (Вербицкая, Казанский, Касевич 2003: 6).

Второй аспект

Ясно (и об этом фактически уже говорилось выше), что если мы предусматриваем для идиом представительство в словаре на правах отдельных статей (единиц), то в словаре появляются слова, которые не являются **фонетическими** словами. Согласно традиции, между тем, любое слово словаря есть одновременно фонетическое слово – имеющее одно ударение. Хотя в тексте не всякая словоформа выступает как фонетическое слово, она может быть лишь **частью** фонетического слова, состоящего из словоформы плюс клитики. (Впрочем, и в традиционной лексико-графической практике находится возможное несоответствие названному правилу, напр., в случаях *рок-музыка* и *дом-музей* : *дома-музея* (причем, данный пример приводится в работе, далекой от фонетических исследований (Мустайоки, Копотев 2004)), как правило, решаемое за счет введения дополнительного, второстепеного ударения). Учет идиом приводит к тому, что текст содержит не только одноударные единицы (фонетические слова, которые могут как совпадать с «лексико-грамматическими» словами, так и включать две-три такие единицы), но и дву- и более ударные сочетания (идиомы). Данная ситуация вызывает проблемы прежде всего с точки зрения восприятия речи. Наличие дву- и более ударных единиц в тексте и перцептивном словаре нарушает, казалось бы, обязательное соответствие «сколько ударений – столько (фонетических) слов», которое лежит в основе первичной сегментации текста при его восприятии[4].

Если же допустить, что слушающий все-таки использует механизм «грубой» акцентной сегментации на слова и при обработке идиом, то из этого следует, что идиомы при восприятии речи должны синтезироваться из словоформ-компонентов на каком-то более позднем этапе (что, по-видимому, подтверждало бы адекватность «опции» с раздельным описанием компонентов идиом). Впрочем, возможен подход, при котором класс идиом оказывается структурно неоднородным с точки зрения процедурной

[3] Таким образом, мы отходим здесь от абсолютизации представлений, согласно которым, например, слово *друг* из местоименного (составного) слова *друг друга* не имеет никакого отношения к слову *друг* из сочетаний *верный друг*, *помощь друга* и т.п. (Венцов, Касевич 1998).

[4] В предыдущих работах авторов было экспериментально показано, впрочем, что в русском языке существует редукция ударения, которая приводит к неадекватному членению текста на слова (в результате чего, например, слова *барбариса* и *штукатурка* не отличаются от словосочетаний *бар Бориса* и *штука турка* соответственно – Касевич, Ягунова 2003).

целостности этих единиц. При наличии возможности синтезирования идиом из словоформ-компонентов на сравнительно позднем этапе поверхностного восприятия единицы некоторого подкласса идиом, реализующиеся с большей просодической цельнооформленностью (просодической редукцией и, прежде всего, редукцией ударения), могут описываться как целостная единица (слово) уже на этапе первичной перцептивной сегментации.

Общий вывод из изложенного заключается, по крайней мере, в том, что проблема идиом выходит далеко за границы той сферы, которой традиционно занимаются лексикографы и стилисты: здесь перекрещиваются многие важнейшие проблемы теории языка и речевой деятельности, равно как и ключевые вопросы прикладной лингвистики, от решения которых зависит разработка речевых технологий.

Литература

Ахманова О.С. Толковый словарь лингвистических терминов. М., 1966.

Венцов А.В., Касевич В.Б. Проблемы восприятия речи. СПб., 1994.

Венцов А.В., Касевич В.Б. Словарь для модели восприятия речи // Вестник СПбГУ. Серия 2. 1998, Вып.3. с.32-39.

Вербицкая Л.А., Казанский Н.Н., Касевич В.Б. Некоторые проблемы создания национального корпуса русского языка // Научно-техническая информация. Серия 2. № 6. М., 2003. с.2-9

Виноградов В.В. Основные понятия фразеологии как лингвистические дисциплины // Виноградов В.В. Избранные труды Т.3. М., 1977.

Касевич В.Б. Элементы общей лингвистики. М., 1977.

Касевич В.Б., Ягунова Е.В. Ударение и фонетическое слово в русском языке // Проблемы социо- и психолингвистики. Вып.3. Пермь, 2003, с.19-25.

Копотев М., Мустайоки А. Принципы создания Хельсинкского аннотированного корпуса русских текстов (ХАНКО) в сетях Интернет // Научно-техническая информация. Серия 2. № 6, с.33-36, М., 2003.

Мустайоки А., Копотев М. К вопросу о статусе эквивалентов слова типа *потому что*, *в зависимости от*, *к сожалению* // Вопросы языкознания, № 3, 2004.

Рогожникова Р.П. Толковый словарь сочетаний, эквивалентных слову. М., 2003.

Телия В.Н. Типы языковых значений: Связанное значение слова в языке. М., 1981.

Телия В.Н. Фразеология // Ярцева В.Н. (Ред.) Лингвистический энциклопедический словарь. М., 2002.

Ярцева В.Н. (Ред.) Лингвистический энциклопедический словарь. М., 2002.