

Сложность русской загадки

Ключевые слова: денотат, субститут, загадка, отгадка, словарь, корпус

Краткая аннотация

Russian riddle is generally a short rhymed text in which the reader is invited to identify an object, animal, natural phenomenon or process described in a mysterious, metaphorical and playful way. Relation between the riddle and its solution is not obvious but could be established using Russian dictionaries where definitions connect words of a riddle with its solution. An example is discussed in the paper.

*Что за песни пели сирены
или каким именем назывался Ахилл,
скрываясь среди женщин, --
уж на что это, кажется, мудрёные вопросы
а какая-то догадка и здесь возможна.
--Сэр Томас Браун*

Семантическая связь текста загадки с отгадкой не должна быть очевидной, иначе разгадывание не составит затруднений и не вызовет интереса. Денотат загадки (собственно отгадка) заменяется в тексте загадки некоторой переменной которая может быть поименована «**Волк** железный, а **хвост** льняной» = иголка с ниткой; выражена личным местоимением «**Я** маленька, удаленька, во дворец вступила, царя разбудила» = блоха; либо пустой переменной «**X** Сквозь землю прошел, красну шапочку нашел» = гриб; и т.д. (Цитирование по Д. Садовников, «Загадки русского народа», С-Петербург, 1876 г.) [1]. Во всех случаях вместо переменной должен быть подставлен денотат = разгадка, причем если переменная поименована, иногда, но далеко не всегда, само ее имя дает намек на разгадку «Висит **груша**, нельзя скушать» = лампочка – лампа имеет грушевидную форму, оба существительных женского рода, неодушевленные. В других случаях денотат не согласуется с субститутутом ни грамматически, ни семантически (**волк+хвост**, **иголка+нитка**).

Существует предположение [2], что загадки не предназначены для отгадывания, просто нужно знать и запомнить отгадку, а метафорическая форма загадки лишь способствует ее запоминанию и позволяет взглянуть на объект с неожиданной, непривычной стороны, что способствует развитию воображения и просто формирует представление о том, что предметы не всегда таковы, какими они видятся. Этот аспект особенно важен для детского развития, ведь загадка в наше время бытует именно как инструмент воспитания, а глубинный, сакральный аспект загадки утрачен.

С другой стороны, текст загадки и отгадка, вероятно, должны иметь определенную семантическую связь, то есть вместо субститута можно было бы подставить любую константу и загадка потеряла бы смысл: «что есть на свете», либо содержала бы явное противоречие: «что больше себя самого». Выявление семантической связи может быть основано на толковых словарях русского языка, в том числе современных, постольку поскольку нами рассматриваются только те загадки, лексика которых и сама отгадка понятны современному человеку, ребенку в том числе и значение лексики которых не изменилось или мало изменилось со времени возникновения загадки.

Рассмотрим загадку «**X** На четырех ногах, а не зверь, с перьями, а не птица» = кровать. Следует отметить, что в загадке указаны ограничения на денотат: ~ (**X** ≡ зверь), ~ (**X** ≡ птица).

Словарь Д.Н.Ушакова [3] дает толкование каждого знаменательного слова этой загадки. Не приводя полностью все толкования, выпишем по цепочке, зная, к чему мы должны прийти, только те, которые могут удовлетворяют указанным ограничениям.

Нога 2. перен. Опора, нижний конец (*мебели*, частей сооружений и механизмов в разных отраслях техники) (спец., редко, чаще *ножка*).

Мебель Предметы комнатной *обстановки*: столы, стулья, шкафы, *диваны*, кресла и т. п.

Кровать Предмет домашней *обстановки*, служащий для сна и представляющий собой длинную раму на *четырёх ножках* с двумя спинками, на которую кладется *постель*.

Перина. Длинный мешок, набитый пухом или мелкими *перьями*, употр. как подстилка на *кровать*.

По словарю установлена связь между двумя отличительными признаками денотат: Part-of(*X*, *ноги*), Part-of(*X*, *перья*).

Для определения, какое из выделенных слов является отгадкой, следует обратиться к тезаурусу, в котором указаны отношения гипонимии, Нуно и меронимии, Part-of: Нуно(*Мебель*, *Кровать*); Part-of(*Кровать*, *Ноги*); Part-of(*Кровать*, *Перина*); Part-of(*Перина*, *Перья*). По транзитивности отношения Part-of имеем: Part-of(*Кровать*, *Перья*), Part-of(*Кровать*, *Ноги*). Отгадка: $X = \{Кровать\}$. Воспользовавшись словарем Т.Ф.Ефремовой [4], в качестве отгадки получим $X = \{Диван\}$, поскольку в дефиницию для *Кровать* не включено эксплицитно слово *Постель*.

Однако, отгадкой к этой же загадке может служить «*Пьяный индеец*» (стоит на 4 ногах и украшен перьями).

Более сложной представляется загадка вида: «Бежит мужик по лесу, зеркало за поясом» В этом случае словари не показывают прямой связи между загадкой и отгадкой, они выражены ассоциативными связями, которые можно отыскать в корпусе [5]

Топор – пояс – блеснуть - зеркало

Ещё сын охотника заткнул за пояс топор,... [М. М. Пришвин. *Кладовая солнца* (1945)]

Владимир отворил дверь и наткнулся на человека, прижавшегося в угол — топор блеснул у него... [А. С. Пушкин. *Дубровский* (1833)]

Синеватая вечерняя мгла стелилась над болотистой равниной, розовыми зеркалами блестели озёрки, и над верхушками маленьких сосен в просвете туч загорелось несколько звезд. [И. А. Ефремов. *Лезвие бритвы* (1959-1963)]

Разгадка связана с загадкой по корпусу. *зеркало* = *топор*, но может быть также найдена связь *мужик* = *ручей* (бежит по лесу и блестит как зеркало).

Неоднозначность отгадки коренится в самой природе загадки, т.е. отгадку все-таки следовало знать, а не находить. Это знание, вероятно, служило для определения отношения свой-чужой в традиционных культурах. В то же время, поиск связей между загадкой и отгадкой позволяет определить меру сложности загадки и расположить последние на некоторой шкале, одним концом которой (1) может служить вокабула + дефиниция в словаре, а другой конец (0) представлен парой логин - пароль в системе идентификации пользователя, причем пароль должен быть сгенерирован автоматически, исключая любую смысловую связь с логином.

Поиск по словарям

С целью определения именно этой меры сложности, вновь обратимся к словарям. В нашем распоряжении имеются электронные (текстовые) версии толковых словарей: Малый Академический Словарь, Словарь Ефремовой, Словарь Ожегова, Словарь Ушакова, Словарь Даля.

В качестве наиболее близкого к лексике периода середины 19 века, к которому относится и сборник Садовникова, мы выбрали Словарь Даля

Суть метода заключается в следующем:

- а. Составляем лемматизированный словник загадок и отгадок – файл Sadov

- b. Выделяем в Словаре Даля те статьи, в дефинициях которых содержатся слова из словника Sadov – файл DahlSadov
- c. Для каждой отгадки строим дерево достижимости T в DahlSadov:
Скажем, отгадка = w_0 . В дефиниции w_0 имеются слова w_{i1} ; записываем эти слова в список T; далее, в дефиниции каждого w_{i1} имеются слова w_{i2} ; дописываем эти слова, кроме уже вошедших в T, в список T; и т.д. На некотором шаге k процесс прекращается, то есть список T перестает расти. В нашем случае k не превышает 5. В процессе построения T для каждого вновь входящего в него слова w_{ik} указываем родительское слово $w_{i(k-1)}$, что дает возможность построить путь из любого слова в корень – w_0 .
- d. Для каждого слова w_i отгадки определяем длину пути d_i по T до отгадки
- Вычисляем C - меру сложности загадки:
 $C = \sum (1/d_i) / n$, $i=\{1 \dots n\}$, где n – общее число знаменательных слов загадки.
- Сортируем загадки по сложности

Рассмотрим в качестве примера загадки с отгадкой = *Горшок*. Всего таких загадок у Садовникова приведено 72, они записаны в Псковской, Новгородской, Олонецкой, Вологодской, Ярославской, Московской, Тульской, Курской, Рязанской, Нижегородской, Самарской, Тамбовской, Пензенской губерниях, Зап. Сибири, Нерчинском крае, то есть по всей Руси великой.

Некоторые загадки включены в состав словарной статьи отгадки:

Был ребенок, не знал пеленок, стар стал, пеленаться стал

В этом случае $C = 1$, поскольку длина пути по словарю от отгадки до каждого слова загадки $d_i = 1$

горшок = быть (1); знать (1); пеленаться (1); пеленка (1); ребенок (1); старый (1); стать (1);

Некоторые отгадки вообще не выводятся из словаря Даля, то есть ни одно из слов загадки не представлено в Словаре Даля, напр. *Замолола безголова (Коловорот)*. В этом случае $C = 0$.

В остальных случаях каждой загадке поставлено в соответствие число C ($0 < C < 1$), напр. *Никто таков, как Иван Русаков! Сядет на конь да поедет в огонь.*

Пути из отгадки к словам загадки по Словарю Даля следующие:

Горшок - высокий - Иван (2); Горшок - конь (1); Горшок - живот - огонь (2); Горшок-баба - поехать (2); Горшок - пестрый - русак (2);

В данном случае мера сложности (простоты) загадки $C = (1/2 + 1/1 + 1/2 + 1/2 + 1/2) / 6 = 0.5$

Проведя такие вычисления для каждой загадки получим шкалу сложности загадки в относительно Словаря Даля.

На Рис. 1 загадки с отгадкой *Горшок* отсортированы по их сложности или, скорее, по возрастанию простоты для отгадывания: по оси X задан номер загадки, по оси Y – мера простоты, C, от 0 до 1- в смысле словаря Даля. То есть, имея в своем распоряжении Словарь Даля, достаточно легко разгадать загадку с величиной C близкой к 1 и трудно разгадать загадку с величиной C, близкой к нулю.

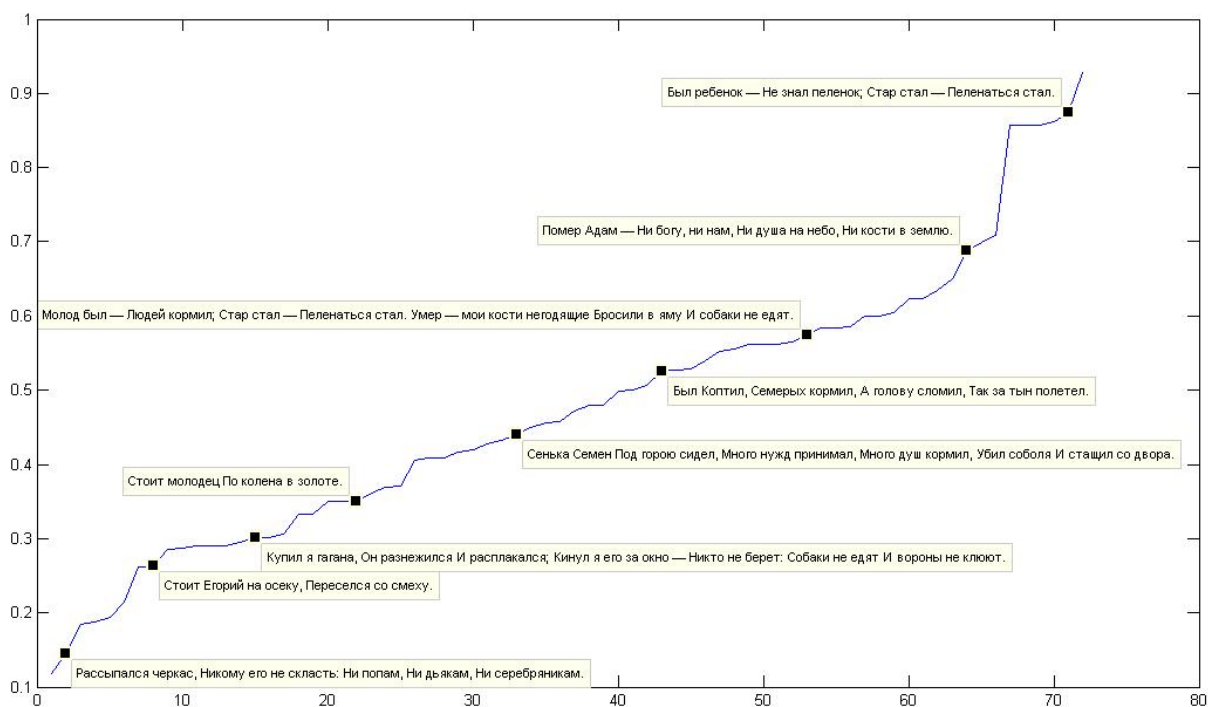


Рис. 1 Шкала сложности (простоты) загадок с отгадкой *Горшок*.

Та же процедура выполнена для словаря Ушакова, для которого построена шкала $S_{\text{Ушаков}}$. В подавляющем большинстве случаев $S_{\text{Ушаков}} < S_{\text{Даль}}$

Объяснением этого может служить тот факт, что в Словаре Даля присутствуют в качестве примеров загадки, чего нет у Ушакова, а также у Даля имеется множество регионализмов, диалектизмов и архаизмов середины 19 века, что характерно и для сборника Садовникова, в то время как Словарь Ушакова отразил русский литературный язык начала 20 века.

Поиск в корпусе

Для определения сложности (простоты) загадки по корпусу текстов использовался Национальный корпус русского языка (НКРЯ) [5]. Запрос к корпусу включает два слова: отгадка + *i*-е знаменательное слово загадки. Задаем расстояние между вхождениями слов в предложение (6). После получения ответа записываем число вхождений пары слов в предложения корпуса, например

Загадка:

Чернец-молодец По колена в золоте. (Горшок)

Запросы и число вхождений (N_i) (на 07.01.2012, позже это число возрастает с ростом НКРЯ)

горшок+чернец (0); горшок+молодец (1); горшок+колена (4); горшок+золото (5);

Вычисляем меру простоты загадки:

$C = \sum_{(i=1...n)} N_i$, где n – общее число знаменательных слов загадки.

Для приведенной загадки $C = 10$. Чем больше величина C , тем загадка проще для разгадывания относительно НКРЯ. На рис. 2 приведена гистограмма числа загадок в зависимости от значения C . Для большинства загадок (более 50) значение C менее 10, например для $C = 2$ число загадок равно 9. Однако имеется 4 загадки со значением $C > 100$. Означает ли тот факт, что в НКРЯ имеется много вхождений пар отгадка+слова загадки, простоту этой загадки? Ответ отрицательный. Например, короткая загадка *Помер Адам — Ни богу, ни нам, Ни душе на небо, Ни кости в землю.* с величиной $C = 115$ не дает практически никакого намека на отгадку (*Горшок*).

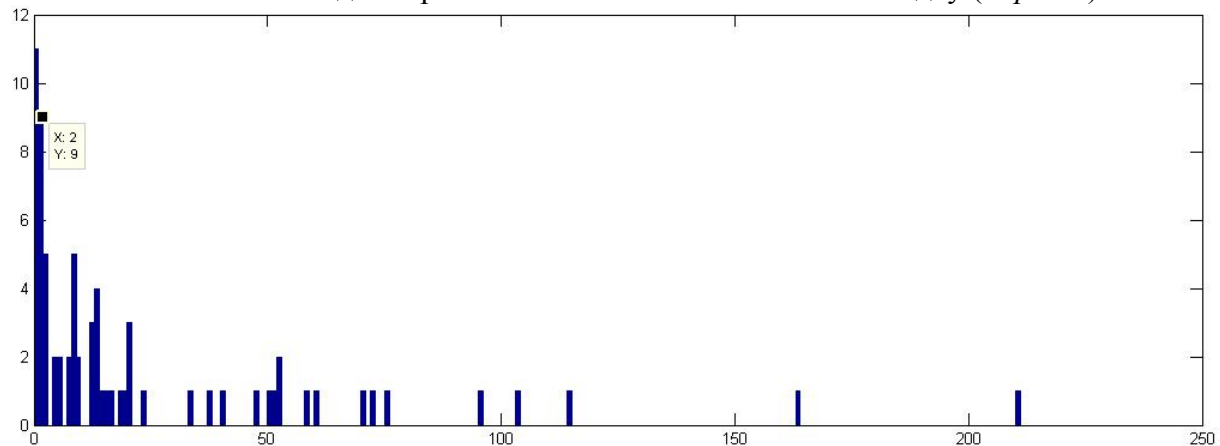


Рис. 2 Гистограмма C

К сожалению, нам не удалось определить разумный нормировочный коэффициент, чтобы привести значение сложности загадки в интервал $(0, 1)$ и обеспечить сравнимость. Нормирование на длину (число знаменательных слов) загадки не дает информативного показателя, поскольку длинная загадка, вероятно, дает больше намеков на отгадку.

Заключение

Сделана попытка формального подхода к установлению соответствия между текстом загадки и отгадкой. Применялся поиск соответствия по словарям и по корпусу текстов. Для словарного поиска определена шкала сложности / простоты загадки относительно данного словаря. Установлено, что мера сложности загадки зависит от используемого словаря, т.е. чем ближе период составления словаря к периоду бытования загадки тем выше возможность отгадать загадку с использованием словаря. Поиск по корпусу не позволил определить разумную меру сложности загадки.

Остаются нерешенными вопросы – как обрабатывать загадку с отгадкой состоящей из 2 или нескольких слов. Также не решена проблема учета отрицания, которое

используется очень часто в тексте загадки. Такие лексемы как не, нет, без теоретически должны ограничивать область поиска, однако как учитывать этот факт при поиске по словарю или по корпусу пока неясно.

Сложная природа и форма загадки предъявляет высокие требования к методам ее обработки, в частности, загадки различной сложности могут быть использованы для тестирования информационно-поисковых систем.

[1] Д. Садовников, «Загадки русского народа», С-Петербург, 1876 г.

[2] С.Я.Сендерович, «Морфология загадки», Из-во Языки славянской культуры, М.2008

[3] Словарь Д.Н.Ушакова [электронная версия, <http://ushakovdictionary.ru/>]

[4] Новый словарь русского языка. Толково-словообразовательный. – М.: Русский язык, 2000

[5] НКРЯ (<http://www.ruscorpora.ru/en/index.html>)