УДК 811.161.1’33:398.21

АВТОМАТИЗИРОВАННЫЙ ПОИСК ЦВЕТА В РУССКИХ СКАЗКАХ

А. В. Рафаева

Московский государственный университет имени М. В. Ломоносова

Поступила в редакцию 18 мая 2015 г.

Аннотация: *статья посвящена проблеме разработки автоматизированного метода нахождения слов с заданной семантикой в тексте сказок с привлечением толковых словарей. Метод рассматривается на примере слов с семантикой цвета. В качестве первоначального источника для отработки метода привлекаются сказки из сборника А. Н. Афанасьева, включенные в состав системы СКАЗКА.*
Ключевые слова: *семантика, словари, сказки, цвет, автоматизированная обработка.*

Abstract: *the article is devoted to the problem of development of automated methods for the search of words with given semantics. Color terms and their semantics were chosen to demonstrate the use of method. The collection of A. N. Afanasiev’s tales from the database of SKAZKA computer system was used as the main source of texts.*
Key words: *semantics, dictionaries, folktales, color, automated proceeding.*

В настоящей статье описывается изучение цвета в русских сказках с помощью автоматизированной системы СКАЗКА [1]. Работа преследовала две цели. Первая – создать, апробировать и описать алгоритм подбора ключевых слов для поиска в сказках материалов, относящихся к некоторой заданной пользователем семантической области. Например, пользователь может захотеть подобрать материал для изучения того, как в сказках проявляют себя животные или сверхъестественные существа, какими профессиями владеют сказочные герои или как представлены в текстах сказок термины родства. Вторая цель – провести компьютерный эксперимент по разрабатываемой методике и описать его результаты.

Для работы был выбран цвет и его представление в русских сказках; одна из причин такого выбора – это сравнительно небольшое количество и частота лексики, обозначающей цвет, в текстах сказок, что позволило сосредоточиться на методе исследования. Эксперимент проводился на материале всех текстов, включенных в систему СКАЗКА, однако обработка и интерпретация полученных результатов в настоящее время проведена только для текстов из сборника А. Н. Афанасьева [2], эти результаты и представлены в статье. Для экспериментов использовались как данные и программное обеспечение, входящие в систему СКАЗКА, так и ряд толковых словарей в электронной форме, доступных на сайтах [3; 4]. Результаты, полученные при работе со словарями в электронной форме, проверялись по толковым словарям [5–8].

Система СКАЗКА включает в себя ряд возможностей для автоматического и автоматизированного анализа текста. В описываемом эксперименте используются частотный словарь всех словоформ, созданный программно по текстам сказок, и программа поиска контекстов для заданной последовательности символов или списка таких последовательностей. Для наглядного представления результатов используется также открытый пакет для работы с графами [9] и программа, разработанная для ручного описания семантических отношений в формате Graphviz. В качестве источника внетекстовой информации предполагается использовать толковые словари. В нашем случае задача сравнительно проста: по словарной статье необходимо узнать, имеет ли хотя бы одно из значений толкуемой лексемы отношение к цвету, после чего произвести проверку по частотному словарю и автоматически найти конкордансы всех лексем, удовлетворяющих заданному требованию.

Иными словами, основная задача эксперимента – разработать метод автоматического поиска лексики, принадлежащей к заданной семантической категории, по словарным статьям выбранных для исследования толковых словарей. Второй задачей является отбор словаря или словарей, в наибольшей степени подходящих для такого поиска.

Работа производилась в несколько этапов.

На первом этапе исследовался метаязык описания, принятый в разных толковых словарях, и определялось, насколько эти описания пригодны для автоматического поиска лексики с заданной семантикой. Для этого делалось следующее:

1. По частотному словарю вручную отбирались словоформы из числа встретившихся в корпусе сказок, которые обозначают цвет или могут иметь отношение к цвету, например, *красный*, *алый*, *белокаменный*, *каурка*. Список может не быть полным, поскольку на

этом этапе самое важное – выявить правила для автоматического поиска лексем с заданной семантикой по словарным статьям выбранных словарей. Уже на этом этапе было очевидно, что большое количество цветов, особенно производные слова (*красно-жёлтый*) и обозначения оттенков (*багровый, алый* и др.) определяются через другие цвета. Это правило применимо ко всем словарям, используемым в эксперименте.

2. Для отобранной лексики в полуавтоматическом режиме создавались цепочки толкований, представляющие отношение *толкуется с помощью* в виде графа. Отношения между отдельными лексемами записывались вручную в явном виде и служили основой для последующего составления правил; графическое же представление служило для наглядного отображения отношений между лексемами и помогало быстрее обнаружить как неполноту, так и избыточность полученных правил.

На втором этапе работы проверялись и уточнялись полученные ранее правила, а также было принято решение о целесообразности использования отдельных толковых словарей в зависимости от задач исследования и вида текстов (в нашем случае – волшебных сказок).

Наконец, на третьем этапе полученные правила и выводы использовались для автоматического отбора текстов (или их фрагментов), отражающих выбранную семантическую область: из словарей извлекался список лексики для поиска в сказочных текстов, проводился поиск текстов и полученные результаты анализировались вручную. Этот этап исследования в настоящее время завершен лишь частично и, по всей вероятности, допускает дополнительную автоматизацию для ускорения и упрощения работы.

При проведении эксперимента по частотному словарю словоформ, составленному для сказок из сборника Афанасьева, вручную было выбрано более 350 словоформ, возможно, выражающих семантику цвета. В список попали словоформы трех основных видов: общеузуальные слова с широкой сочетаемостью (*алый, белый, синий, зеленый* и т.п.), слова, служащие для обозначения масти животных или изредка цвета волос людей (*вороной, бурый, каурый, седой, рыжий*) и производные слова, в том числе употребляемые преимущественно в сказочном узусе (*аленький, бледная-бледная, белёшенька, белокаменный, вызолотить, златорогий, каурка*). Для дальнейшей работы были выделены следующие лексемы: *аленький, алый, красный, золотой, багровый, каурый, рыжий, желтый.* Для сравнения привлекались также словарные статьи для слов *синий, зеленый, златорогий, сизокрылый* и *голубой*, но в результирующие графы они не включены.

Графы строились следующим образом: для каждой лексемы из списка отдельно в каждом из словарей, использованных в эксперименте, просматривались все толкования и вручную определялось, каким именно образом это толкование построено. Таким образом определялись пары лексем, связанных отношением *толкуется с помощью* и сходными с ними, и составлялся список правил, по которым построены подобные толкования. Лексемы, выделенные при толковании первоначального списка, становились материалом для следующего этапа работы, и цикл повторялся до тех пор, пока цепочка толкований не обрывалась, не замыкалась в кольцо или же хотя бы одна лексема не толковалась явным образом с помощью слова *цвет* (подробнее про основные принципы составления цепочек толкований см. [10; 11]). Полнота выделенного подмножества толкований проверялась по результирующему графу, который в случае необходимости дополнялся новыми связями и лексемами. Например, в словаре В. И. Даля [5] *алый* толкуется как *ярко-красный* или *светло-красный* (рис. 1); толкования для этих составных слов отсутствуют (т.е. для продолжения работы приходится вводить отношение *производное от*, в нашем примере это будут лексемы, производные от слова *красный*), а одно из значений лексемы *красный* имеет следующее толкование: «*Красный* – по цвету: рудой, алый, чермный, червленой : кирпичный, малиновый, огневой и пр. разных оттенков и густоты» (приводится по [3]). Очевидно, однако, что список оттенков красного цвета также может пригодиться в дальнейшей работе, потому на этом анализ словарных статей не останавливается, и для каждого из оттенков просматриваются соответствующие словарные статьи, если они представлены в словаре. В противном случае список оттенков, извлеченный из толкования, запоминается для включения в число ключевых слов при полнотекстовом поиске (рис. 2, 3).

Приведем графы, полученные для словарей [5–7]. К сожалению, все доступные электронные версии словаря [8], которые удалось проверить, слишком низкого качества и не могут быть использованы в дальнейшей работе. Например, в них отсутствует словарная статья для лексемы *жёлтый* (как в форме *желтый*, так и в форме *жёлтый*).

Сделаем некоторые предварительные выводы. Все из участвующих в эксперименте словарей неполны с точки зрения нашего корпуса; например, ни в одном из них нет слова *аленький*, сравнительно часто встречающегося в сказках. Поэтому для того, чтобы полностью отобрать всю лексику, имеющую отношение к заданной тематике, потребуется дополнительная ручная работа с частотным словарем словоформ, составленном по базам данных сказок, хотя эта работа будет сильно облегчена по сравнению с поиском необходимой лексики полностью вручную.

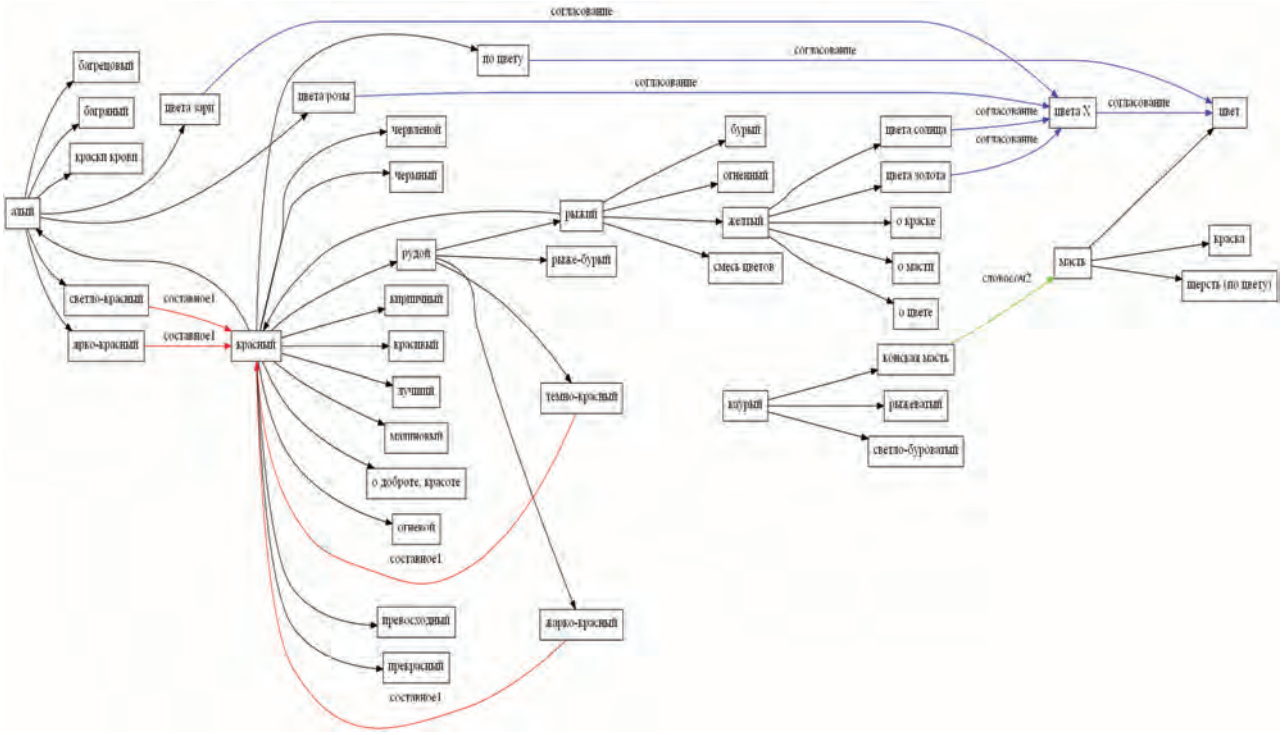

Рис. 1. Связи между цветами по словарю Даля

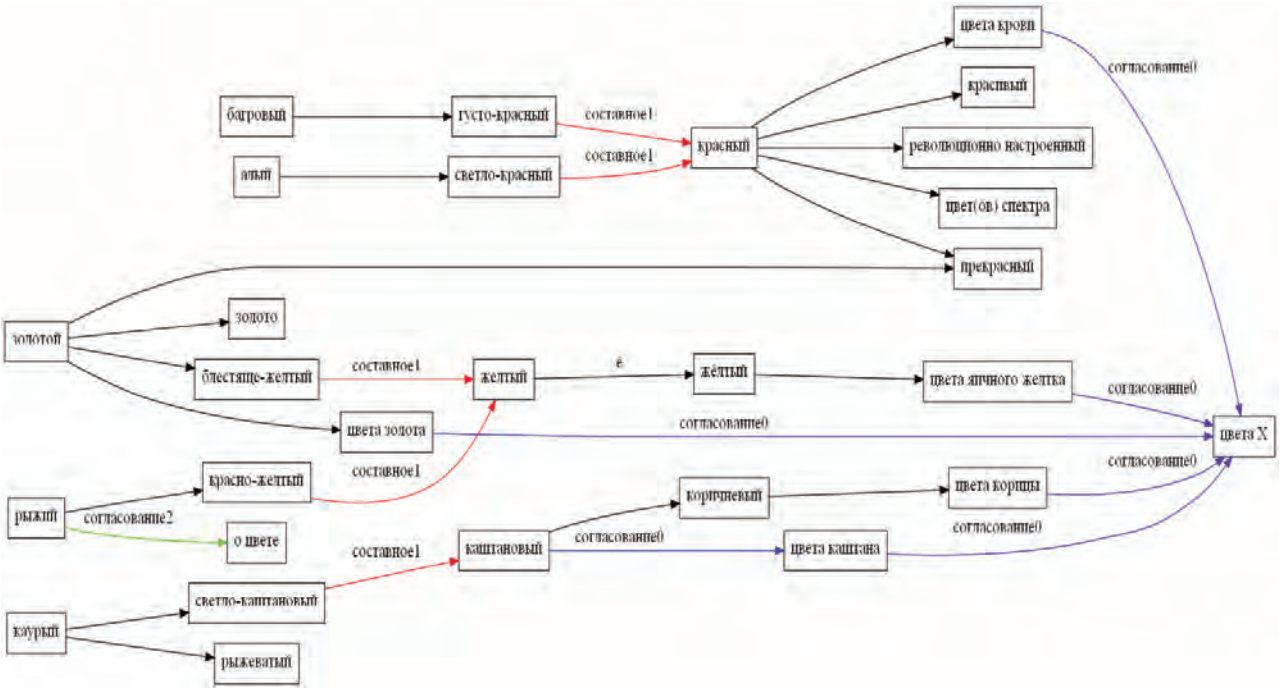

Рис. 2. Связи между цветами по МАС

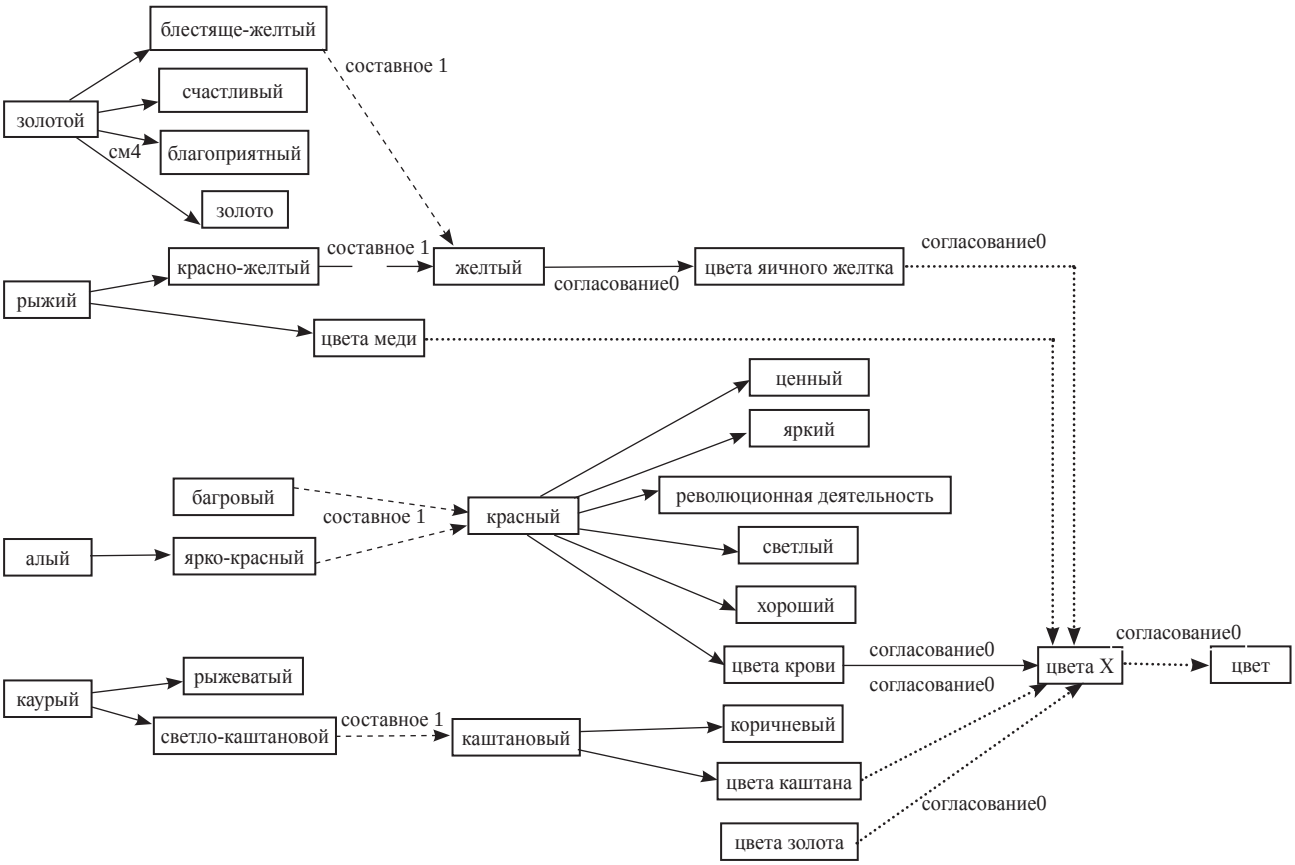

Рис. 3. Связи между цветами по словарю Ожегова

Даже по виду приведенных графов можно предположить, что язык метаописания и правила толкования в МАС формализованы наиболее строго, а число правил минимально. В дальнейшем это предположение подтвердилось. Самый непоследовательный и наименее формализованный метаязык описания, как можно было предположить заранее, представлен в словаре В. И. Даля. Однако задача автоматического поиска по словарным статьям этого словаря облегчается тем, что во многих статьях одновременно используется сразу несколько правил, а программа включает в список все слова, в толковании хотя бы одного из значений которого использовано по крайней мере одно правило из числа заданных. Кроме того, в словаре Даля собрано большое количество диалектной и устаревшей лексики. Наконец, наименее удобным для автоматической обработки в предложенном эксперименте оказался словарь С. И. Ожегова.

Приведем правила, общие для всех трех словарей.

1. Цвета Х, где Х – существительное или именная группа (*цвета крови*, *цвета яичного желтка*, *цвета цветов василька*). Это правило широко используется, но его можно применять только с последующей ручной обработкой или при наличии синтаксического анализа, хотя бы частичного.

2. То же, что Х, где Х – обозначение цвета (*лазоревый – то же, что лазурный*). Правило полностью автоматизируется.

3. Составные номинации (*светло-красный*, *тёмно-синий*). Правило легко автоматизируется.

4. Перечисления X1, X2, … XN, где некоторые Xi – обозначения цветов. Правило легко автоматизируется.

5. (МАС) Словообразование: златорогий, сизокрылый и т.п. Частично автоматизируется, однако может потребовать дополнительной обработки.

Словарь В. И. Даля дает ряд дополнительных правил, однако они отчасти избыточны, т.е. словарная статья, как правило, будет соответствовать сразу нескольким признакам (см. приведенный выше пример словарной статьи для слова *красный*). Также есть ряд правил, отдельных для каждого из словарей, в которых цвета толкуются через метаслова *оттенок*, *окраска*, *спектр*, *масть* (животного).

Более общий алгоритм выделения правил для нахождения ключевых слов по толковым словарям будет выглядеть следующим образом:

• по частотному словарю к корпусу текстов выделить достаточно представительный набор лексики, отвечающей заданному критерию;

• выбрать подмножество единиц для проверки по словарным статьям и выделения правил;

• составить граф, отражающий правила и связи между единицами наглядно;

• описать сами правила, особенно обращая внимание на легко автоматизируемые и наиболее употребительные;

• найти исключения, слова, не представленные в толковых словарях. Если возможно, формализовать их отличия. Так, для сказочной лексики это часто будут производные слова, в том числе образованные с помощью повтора или уменьшительных суффиксов. Однако в некоторых случаях составители словарей отступают от предложенной ими самими неявной схемы описания и предлагают толкования, не подходящие ни под одно из выделенных правил. Поэтому в том случае, когда лексику из заданной семантической области необходимо найти полностью, а также в случае обращения к сказкам, собиратели которых стремились передать на письме особенности устной речи, потребуется проверка по частотному словарю словоформ, составленному для корпуса текстов;

• в полуавтоматическом режиме (с дополнительной ручной работой по частотному словарю словоформ, если необходимо) составить список ключевых слов для полнотекстового поиска.

Описанный алгоритм поиска выдал большое количество контекстов, в которых упоминается цвет. Работа над полученными результатами в настоящее время продолжается, поэтому ниже они будут описаны очень кратко и прежде всего с точки зрения того, какие еще средства автоматической обработки должны быть созданы или добавлены дополнительно.

Прежде всего следует заметить, что сборник Афанасьева отражает сказочные явления (в том числе касающиеся цвета) недостаточно полно. Например, *песок* в этом источнике всегда *желтый*, в то время как в других источниках изредка встречаются также *белый*, *красный* и *черный песок*. То есть для более полного изучения семантики цвета в сказках, несомненно, потребуется использование других сказочных сборников. Однако если сказки из сборника Афанасьева записаны литературным языком или близко к нему, что позволяет пользоваться при анализе словарями и лингвистическими программами, то ряд других сборников отражает на письме особенности диалектного произношения, и это сильно затрудняет задачу автоматизированного анализа. Возможным выходом из этого затруднения была бы либо полная разметка текста, либо использование более грубых статистических методов.

Число цветов и прилагательных цвета, используемых в сказках, сравнительно невелико. Наиболее частыми являются *золотой* и *красный* (811 и 601 употребление, соответственно), однако оба этих прилагательных используются не только для обозначения цвета чего-либо, но и в других значениях: *золотой* – для обозначения материала (*золотой перстень*), *красный* – в значении красивый.

В ряде случаев невозможно определить, присутствует ли семантика цвета в словосочетании. Например, в устойчивых словосочетаниях *красна девица*, *красно солнышко* прилагательное *красный* используется в значении *красивый*, однако иногда по тексту бывает сложно установить однозначно, означает ли словосочетание *красное платье* одежду красного цвета или красивое платье (ср. описание девушки: *краше цвета алого, белей снегу белого*).

Приведем список остальных обозначений цвета в порядке убывания частоты упоминания: *белый*, *серебряный* (с той же оговоркой, что и для прилагательного *золотой*), *серый*, *синий*, *зеленый*, *черный*, *сивый*, *вороной*, *каурый*, *бурый*, *алый*, *седой*, *сизый*, *рыжий*, *голубой*, *желтый*, *червонный*, *лазоревый*, *изумрудный*. Под вопросом использование *багрового*, словосочетание *вышничек багровый* выявлено единожды при перепечатке лубочной сказки и обозначено как сомнительное.

По сочетаемости прилагательные цвета можно разделить на следующие группы: прилагательные цвета и производные слова с широкой сочетаемостью (*золотой*, *красный*, *белый*, *серебряный*, *синий*, *зелёный*, *чёрный*, *серый*), слова, обозначающие масть животного (*сивый*, *вороной*, *каурый*, *бурый*, *сизый*), слова, обозначающие масть животного или цвет волос человека (*рыжий*, *седой*, *чернявый*), и, наконец, слова, служащие для описания одежды и обозначения красивых предметов, часто обладающих чудесными свойствами (*голубой*, *алый*, *лазоревый*, *багровый (?)*, *червонный (?)*, *изумрудный*).

Отдельный интерес представляет анализ устойчивых сочетаний, например: *синее море*, *зеленые луга* и т.п. Эта работа сейчас ведется, предполагается ее дальнейшая автоматизация.

Цвет, точнее, указание на яркость и красочность чего-либо (чаще всего одежды, дворцов и чудесных построек) или пестроту окраски животных и птиц может передаваться в сказках не только при помощи прилагательных цвета. Иногда цвет таких объектов не указывается явно, а задается с помощью определений *цветной* (чаще всего в сочетаниях *цветное платье*, *цветные луга*), *пестрый* (*пестрые крылышки*, *пестрый пес*, *пестрая оса*), реже некоторых других.

В других случаях цвет передается через обозначение материала, который одновременно служит и для отражения «красоты», красочности открывающейся картины, чаще всего это *медный*, *серебряный* и *золотой* (материал и цвет), приведем более редкий пример: герой *приходит к такому дворцу, что и господи боже мой! — так и горит в бриллиантах и самоцветных каменьях* ([2, Т. I, с. 191]). Такие случаи

требуют отдельного изучения. Еще более редкий способ задания цвета в сказках – с помощью сравнений (*краснее солнца, яснее месяца и белее снегу*).

Представляет несомненный интерес сопоставление семантики цвета в сказках с общеузуальной семантикой. Эту работу имеет смысл проводить после того, как будет подробнее рассмотрен сказочный материал и сделаны хотя бы предварительные выводы. Таким образом, использование программного обеспечения для анализа цвета в русских сказках позволило получить более полный материал, однако поставило новые исследовательские задачи, в том числе и разработки дополнительных программных средств и алгоритмов анализа.

ЛИТЕРАТУРА

1. *Рафаева А. В.* Компьютер – Слово – Фольклор / А. В. Рафаева. – М., 2014. – 280 с.

2. Народные русские сказки А. Н. Афанасьева : в 3 т. / подг. Л. Г. Бараг, Н. В. Новиков ; отв. ред. Э. В. Померанцева, К. В. Чистов. – М., 1984–1985.

3. Reword : бесплатная программа-словарь. – Режим доступа: http://reword.org/online

4. Словари Онлайн. – Режим доступа: http://slovarionline.ru

5. *Даль В. И.* Толковый словарь живого великорусского языка : в 4 т. / В. И. Даль. – 4-е изд., стер. – М., 2007.

6. Словарь русского языка : в 4-х т. / РАН; Ин-т лингвист. исследований ; под ред. А. П. Евгеньевой. – 4-е изд., стер. – М., 1999. – Режим доступа: http://feb-web.ru/feb/ushakov/ush-abc/default.asp

7. *Ожегов С. И.* Толковый словарь русского языка : 80 000 слов и фразеологических выражений / С. И. Ожегов, Н. Ю. Шведова. – 4-е изд., М., 1997. – 944 с.

8. Толковый словарь русского языка : в 4 т. / под ред. Д. Н. Ушакова. – М., 1935–1940. – Режим доступа: http://feb-web.ru/feb/ushakov/ush-abc/default.asp

9. Graphviz – Graph Visualization Software. Envisioning connections. – Mode of access: www.graphviz.org

10. *Кретов А. А.* К созданию компьютерной системы семантической классификации лексики / А. А. Кретов, А. В. Рафаева // Проблемы лингвистической прогностики : сб. науч. трудов. – Вып. 4. – Воронеж, 2007. – С. 53–69.

11. *Кретов А.А.* Программа семантической классификации лексики – ПроСеКа : теоретические и прикладные аспекты / А. А. Кретов, А. В. Рафаева // Компьютерная лингвистика и интеллектуальные технологии : по материалам ежегод. Междунар. конф. «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). – Вып. 8 (15). – М. : РГГУ, 2009. – С. 230–235.

Московский государственный университет имени М. В. Ломоносова

Рафаева А. В., научный сотрудник лаборатории автоматизированных лексикографических систем

E-mail: anna_raf@rambler.ru

Тел.: 8 (495) 939-23-57

Moscow State University named after M. V. Lomonosov

Rafaeva A. V., Research Assistant of the Computational Lexicography Laboratory

E-mail: anna_raf@rambler.ru

Tel.: 8 (495) 939-23-57