ВИЗУАЛИЗАЦИЯ ДАННЫХ ДЛЯ КАТАЛОГА РУССКИХ ЛЕКСИЧЕСКИХ КОНСТРУКЦИЙ (НА МАТЕРИАЛЕ НКРЯ)

Ляшевская О. Н. (olesar@gmail.com)

НИУ Высшая школа экономики, Москва, Россия; Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Митрофанова О. А. (alkonost-om@yandex.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Паничева П. В. (ppolin86@gmail.com)

EPAM Systems, Россия

Доклад отражает новые результаты, полученные в ходе совместного проекта кафедры математической лингвистики СПбГУ с разработчиками Национального корпуса русского языка (НКРЯ, http://ruscorpora.ru). Цель исследования — разработка технологии автоматического распознавания в тексте конструкций, связанных с той или иной лексической единицей, и применение этой технологии в создании каталога русских лексических конструкций. Выделение конструкций предполагает использование потенциала многоплановой лингвистической разметки НКРЯ (прежде всего, лексико-семантической). В докладе обсуждается использование модуля визуализации данных для уточнения информации о конструкциях, реализующих значения исследуемых слов. Модуль предназначен для лексикографов и исследователей лексики и грамматики русского языка.

Ключевые слова: именные конструкции, сочетаемость, НКРЯ, лексико-семантическая разметка, лексико-грамматическая разметка, визуализация данных

DATA VISUALIZATION FOR BUILDING THE CATALOGUE OF RUSSIAN LEXICAL CONSTRUCTIONS (BASED ON RNC)

Lyashevskaya O. N. (olesar@gmail.com)
NRU Higher School of Economics, Moscow, Russia;
Vinogradov Institute of Russian Language RAS, Moscow, Russia

Mitrofanova O. A. (alkonost-om@yandex.ru)
Saint-Petersburg State University, St. Petersburg Russia

Panicheva P. V. (ppolin86@gmail.com)
EPAM Systems, Russia

Our research aims at automatic identification of constructions associated with particular lexical items and its subsequent use in building the catalogue of Russian lexical constructions. The study is based on the data extracted from the Russian National Corpus (RNC, http://ruscorpora.ru). The main accent is made on extensive use of morphological and lexico-semantic data drawn from the multi-level corpus annotation. Lexical constructions are regarded as the most frequent combinations of a target word and corpus tags which regularly occur within a certain left and/or right context and mark a given meaning of a target word. We focus on nominal constructions with target lexemes that refer to speech acts, emotions, and instruments. The toolkit that processes corpus samples and learns up the constructions is described. We provide analysis for the structure and content of extracted constructions (e. g. r:ord der:num t:ord r:qual|*pervyj 'first'* + LJUBOV' 'love'; LJUBOV' 'love' + PR|*s 'from'* + ANUM m sg gen|*pervyj 'first'* + S f inan sg gen|*vzgljad 'sight'* = *love at first sight*). As regards their structure, constructions may be considered as n-grams (n is 2 to 5). The representation of constructions is bipartite as they may combine either morphological and lemma tags or lexical-semantic and lemma tags. We discuss the use of visualization module PATTERN.GRAPH that represents the inner structure of extracted constructions.

Key words: nominal constructions, word co-occurrence, patterns, Russian National Corpus, lexico-semantic annotation, lexico-grammatic annotation, data vizualization

1. Введение

Данная статья продолжает цикл публикаций, посвященных автоматическому выделению лексических конструкций в контекстах НКРЯ (см., например, [Lashevskaya et al. 2012, Mitrofanova et al. 2012]). Цель проекта — предложить основанную на статистических методах технологию автоматического распознавания типичных конструкций, связанных с той или иной лексической единицей.

Современные корпуса и веб-архивы дают возможность собрать статистику о поведении лексической единицы в контексте, составить портрет контекстного окружения слова (Behavioural Profile, [Divjak, Gries 2009]). Дистрибутивная гипотеза [Firth 1957/1968, Sahlgren 2008] предполагает, что различающиеся группы контекстов отражают употребления слова в разных значениях. Например, резкая мена контекстного окружения в корпусе нового времени может обозначать, что у слова появилось новое значение.

Традиционно контекстный профиль представляют как наборы n-грамов (обычно 1…4 словоформ, примыкающих к ключевому слову справа и слева), которые сгруппированы по определенным признакам. На *n*-грамах строятся векторные модели, с которыми удобно работать машине, однако человеческому глазу как векторные таблицы, так и сами списки *n*-грамов не слишком удобны. Секрет в том, что человеку свойственно видеть вместо множеств — структуру: вместо наборов *n*-грамов в словарях и грамматиках содержатся указания на типичные признаки контекста: например, что лексема употребляется в переходной конструкции, управляет творительным падежом, присоединяет тот или иной предлог; приводится наиболее показательная лексическая сочетаемость. Тем самым, какие-то повторяющиеся в *n*-грамах признаки признаются важными, а другие элементы, случайные, отметаются.

Задача автоматического распознавания структуры контекстов призвана перебросить мост между корпусной выборкой, на которой строятся *n*-грамы, и словарем/грамматикой. Адресат модуля визуализации корпусных данных, о котором пойдет речь в статье — лингвист-лексикограф или исследователь лексики, морфологии и синтаксиса. Идея состоит в том, чтобы электронный помощник кластеризовал корпусные примеры на употребление той или иной лексемы и выделял повторяющиеся в контекстном окружении паттерны.

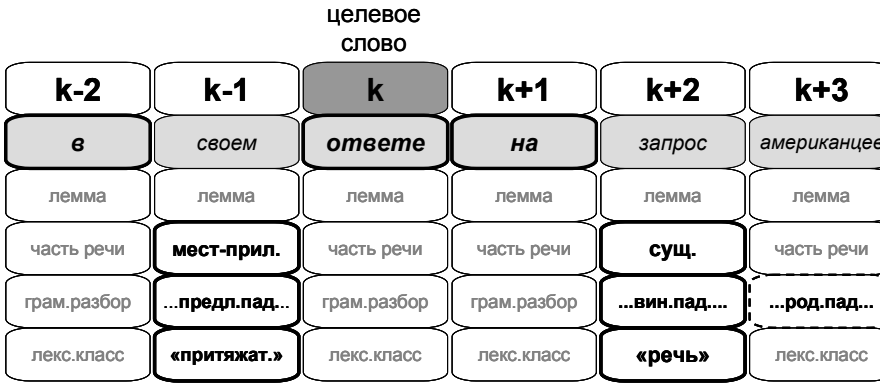

Рис. 1. *N*-грам *в своем ответе на запрос американцев* с разметкой разных уровней

Для выделения повторяющихся признаков можно воспользоваться корпусным арсеналом, чтобы за каждым элементом n-грама стояла разметка разных уровней (часть речи, синтаксическая группа и т. п., для русского языка — лемма

и словоизменительные грамматические характеристики). Тогда *n*-грам *в своем ответе на запрос американцев* (см. Рис. 1) попадет в кластер с повторяющимися элементами, среди которых будут предлоги *в* и *на,* притяжательное прилагательное в предл. падеже, имя из класса «речь» в вин. падеже и слово в род. падеже (элементы расположены в определенном порядке). Более крупный кластер будет включать цепочку *в* + *ответе* + *на* + «речь». Использование разметки разного уровня даст более мощный инструмент, нежели обычные кластеры *n*-грамов словоформ (как в корпусах М. Дейвиса, http://corpus.byu.edu). Он также будет более гибким, нежели Sketch Engine (http://www.sketchengine.co.uk), т. к. набор грамматических паттернов в нем не будет задан заранее.

Наша гипотеза состоит в том, что выделяемые последовательности должны интерпретироваться как лингвистически значимые законченные лексические конструкции (центр конструкции — искомое целевое слово). Вместе с тем, следует иметь в виду, что пользователи могут быть заинтересованы в получении разных конструкций — разной длины, разной степени абстрактности и т. п. Инструмент должен показывать, как изменится паттерн при переходе от более дробных к более крупным кластерам, можно ли изменить расстояние между элементами и что ожидается во «вставке» и т. п. Нужно также предусмотреть взаимодействие конструкции с элементами, традиционно в нее не включаемыми — например, как изменится паттерн, если за целевым словом *ответ* будет следовать частица *же*.

Тема динамической визуализации данных, к сожалению, пока еще редко поднимается в корпусной лингвистике — особенно если речь идет о пользователях, не искушенных в квантитативной лингвистике и в работе со статистическими программами. В этой статье описаны пилотные эксперименты по визуальному представлению именных конструкции, и пока еще далеко не все задачи решены. Однако, мы надеемся, что проблематика статьи вдохновит разработчиков на создание разнообразных визуализаторов корпусных данных в помощь лингвистам.

2. Теоретическая база исследования

В центре исследования находятся именные конструкции — прежде всего, те, которые строятся вокруг имен существительных. Исследовались как конструкции отдельных лексем, так и конструкции, свойственные целым лексико-семантическим группам: обозначениям речевых действий (*дискуссия, комплимент, обращение, обсуждение, ответ* и т. д.), названиям эмоций (*апатия, благодарность, грусть, гнев, любовь* и т. д.), именам инструментов (*бритва, веник, весло, карандаш, коса* и т. д.).

Лексические конструкции — это наблюдаемые в речи последовательности лексических единиц, из которых одно (или несколько) — лексическая константа, а другие — переменные [Fillmore 1988a]. Предполагается, что слово в определенном значении способно структурно организовывать контекст вокруг себя — то есть характеризуется набором лексических конструкций,

которые строятся вокруг нее. Тем самым, основная функция конструкции — фиксировать регулярную сочетаемость целевого слова в определенном его лексическом значении (наполнение слотов ассоциируется с семантикой целевого слова).

Согласно идеологии Грамматики конструкций [Fillmore 1988b, Goldberg 1995, 2006, Tomasello 2003], лексическая конструкция, как и другие виды конструкций, обладает единством формы и значения. Форма лексической конструкции задается, с одной стороны, очевидно, лексически фиксированными единицами, а с другой стороны — ограничениями на заполнение переменных слотов: морфологическими, синтаксическими, лексико-семантическими. Форма конструкции может предусматривать и грамматические ограничения на форму ключевого слова — лексического центра конструкции. Конструкция может реализоваться в виде синтагмы: простого или сложного словосочетания, которое может, например, реализовать рамку валентностей целевого слова или даже выходить за ее пределы.

С точки зрения структурной организации, конструкция — это комбинация целевого слова и слотов, заполняемых регулярными контекстными соседями, среди которых могут быть леммы, грамматические (морфологические и синтаксические), лексико-семантические и т. п. признаки. Точнее говоря, по своей природе, конструкция — это абстрактный шаблон, предполагающий лексикализацию, т. е. различные реализации в виде комбинаций лемм/словоформ, ср. V|*дать, найти, предложить…* ОТВЕТ + PR|*на* + speech r:abstr|*вопрос*, r:qual|*простой, неоднозначный…* + ОТВЕТ, ОТВЕТ + t:hum r:concr|*академикам, мудрецам, отцу…*

Значение конструкции характеризуется большей или меньшей устойчивостью, варьирующей от регулярной свободной сочетаемости до высокой идиоматичности. Значение конструкции, как правило, некомпозиционно, то есть не выводится из значения составляющих элементов (в особенности если рассматривается абстрактный шаблон — комбинация целевого слова и лексико-семантических тегов классов). Лексикализованные конструкции могут удовлетворять принципу композиционности, если в них реализуется типовая свободная сочетаемость. Некомпозиционные сочетания (фраземы), в которых лексически фиксированы все элементы (ср. *любовь с первого взгляда*), также входят в фонд лексических конструкций, наряду с более свободными шаблонами, где ограничения на элементы задаются признаками типа «глагол», «инфинитив», «предлог *на* + предложный падеж».

Конструкция, понимаемая таким образом, это многоярусная структура, призванная компактно и в достаточной мере полно описать сочетаемостные возможности целевого слова, ассоциированные с его лексическим значением, и задать сочетаемость не только в терминах лемм/словоформ, но и с точки зрения грамматических и лексико-семантических классов. Данный взгляд на конструкции отражает идею взаимосвязи и взаимопроникновения различных уровней языка (от фонетического/графического до лексического) и позволяет рассмотреть языковые выражения не в их проекции на один из множества уровней (как представлялось бы с точки зрения модульного подхода), а как многоярусные структуры.

Подводя итог сказанному выше, можно заметить, что наше определение конструкции не противоречит традиционному, однако несколько выходит за его рамки. Принятое в нашем исследовании понимание конструкции позволяет, в отличие от метода *n*-грам, относиться избирательно к сочетаемостным возможностям целевых слов, учитывать тенденции в сочетаемости целевого слова и его соседей в контексте, описывать как лексическую сочетаемость, так и сочетаемость на уровне классов, не только устойчивые, но и свободные сочетания, важным образом отражающие типовое употребление слова в тексте.

Извлечение конструкций базируется на автоматической обработке множества корпусных контекстов, в которых употребляется целевое слово. Контекстные цепочки разбиваются на группы, идентифицируются типовые паттерны и, соответственно, выделяются признаки, обобщающие свойства элементов-соседей. Тем самым, компьютерная система «выучивает» конструкции по принципу генерализации свойств контекстного окружения (ср. гипотезу о генерализации конструкций при усвоении языка детьми [Tomasello 2003]). Нельзя не заметить, что на подобных же основаниях (у элементов с общим значением будет сходное контекстное окружение) действуют системы разрешения лексической неоднозначности (WSD), разметки семантических ролей (Semantic Role Labelling) и многие другие модули автоматической обработки текста. Однако в данном случае речь идет именно об экспликации информации в виде ограничений на элементы контекстных последовательностей.

Далее в статье мы опишем подходы к автоматическому выделению лексических конструкций, опишем эксперименты с выделением конструкций для имен эмоций, речи и инструментов и представим модуль визуализации структуры и наполнения конструкций.

3. Методика автоматического выделения конструкций в выборках НКРЯ

Итак, формально под конструкцией в нашем проекте понимается регулярная комбинация целевого слова (лексической константы) и различных тегов контекстного окружения, присутствующих в многоярусной разметке корпуса (см., в частности, [Lashevskaya et al. 2012, Mitrofanova et al. 2012]). В качестве основного лингвистического ресурса задействован Национальный корпус русского языка (НКРЯ), отличающийся богатством текстового наполнения, а также детальностью и многоплановостью лингвистической разметки. Акцент делается на использование в обучении корпусной разметки — на уровне лемм, частей речи, грамматических признаков, лексико-семантических признаков.

При выделении конструкций учитываются теги лемм, лексико-семантические и морфологические теги (*lex, sem, gr*). В этой связи, как конструкции нами рассматриваются, например, следующие сочетания целевых слов и элементов их контекстного окружения:

(1) ОТВЕТ + PR|*на* + t:speech r:abstr|*приветствие, вопрос, высказывание, рапорт, реплика*

(2) V pf tran inf act|*найти, дать* + A m sg acc inan plen|*простой, однозначный* + ответ + PR|*на* +S m inan sg acc|*вопрос*

(3) r:ord der:num t:ord r:qual|*первый* + ЛЮБОВЬ

(4) ЛЮБОВЬ + PR|*с* + ANUM m sg gen|*первый* + S f inan sg gen|*взгляд*

Семейства конструкций для отдельного слова ассоциируются с его значением, например:

(5) ДИСКУССИЯ + PR|*о, по, на*
r:qual|*горячий, долгий, жесткий, серьезнейший, старый, широкий* + дискуссия
ДИСКУССИЯ + PR|*о* + r:abstr|*вред, ценность, целесообразность, красота*
ДИСКУССИЯ + PR|*на* +t:pers|*тема*
V|*начать,организовать* + ДИСКУССИЯ
V + r:qual + ДИСКУССИЯ + PR + t:pers + r:abstr

В случае многозначности целевых слов каждое из их значений можно охарактеризовать определенным набором конструкций, следовательно, исследование семейств конструкций позволяет разграничивать (в том числе и автоматически) значения многозначного слова.

4. Эксперименты по автоматическому выделению конструкций в выборках НКРЯ

4.1. Инструмент автоматического выделения конструкций

Существует ряд проектов, в которых особое внимание уделяется формализации лексико-синтаксических связей единиц текста, например, PropBank (http://verbs.colorado.edu/~mpalmer/projects/ace.html), NomBank (http://nlp.cs.nyu.edu/meyers/NomBank.html), FrameNet (https://framenet.icsi.berkeley.edu/fndrupal/), DeepDict (http://gramtrans.com/deepdict/), Sketch Engine (http://www.sketchengine.co.uk/), StringNet (http://nav3.stringnet.org/) и т. д. Данные ресурсы дают разноплановую информацию о сочетаемости лексических единиц, при этом форма представления результирующих данных, как правило, табличная. Исключение составляют PropBank и NomBank, где важнее всего оказывается семантико-синтаксическая разметка контекстов.

Для представления данных о конструкциях в рамках нашего проекта был создан специализированный модуль на языке Perl (разработчик С. В. Романов), где используются некоторые стандартные средства для обработки контекстных выборок с многоярусной лингвистической разметкой и для эффективной

выдачи данных (в частности, XML::LibXML, YAML, Log::Log4perl). Важнейший компонент нашего модуля — пакет Algorithm::Combinatorics, с помощью которого производится выявление частотных комбинаций тегов в контекстах для целевых слов.

На вход программы подается файл с выборкой контекстов с целевым словом, для которого требуется выявить конструкции. Затем пользователь определяет такие параметры обработки данных, как типы тегов, учитываемых при выделении конструкций (*lex, sem, gr*), ширина контекстного окна, в пределах которого ведется поиск частотных комбинаций тегов (от −5 до +5), а также число конструкций, попадающих в выдачу (от 1 до 50).

4.2. Анализ результатов работы инструмента автоматического выделения конструкций

Файл с результатами работы инструмента автоматического выделения конструкций содержит наиболее частотные сочетания целевого слова и различных тегов контекстного окружения (*lex, sem, gr*). В зависимости от назначенной ширины контекстного окна в выдачу попадают комбинации тегов в виде пар, троек, четверок, пятерок и т. д. Например, из троек в выдаче присутствуют частотные конструкции, организованные по схемам *sem+sem+sem, sem+sem+gr, sem+sem+lex, lex+sem+sem, gr+sem+sem, lex+sem+lex, lex+sem+gr, gr+sem+gr, gr+sem+lex* и т. д. Например, в случае конструкции *в азарте игры* мы получаем примерно следующие комбинации тегов в выдаче:

(6) *gr+lex+sem* PR|*в* + АЗАРТ + der:v r:abstr der:s|*игра*
gr+sem+gr PR|*в* + t:psych r:abstr + S f inan pl gen|*игра*
gr+sem+sem PR|*в* + t:psych r:abstr + der:v r:abstr der:s|*игра*
lex+sem+gr *в* + t:psych r:abstr + S f inan pl gen|*игра*
gr+gr+sem PR|*в* + S m inan sg loc + der:v r:abstr der:s|*игра*
gr+gr+gr PR|*в* + S m inan sg loc + S f inan pl gen|*игра*

и т. д.

В настоящий момент мы можем получать конструкции с двухслойной структурой, т. е. компоненты конструкции могут одновременно характеризоваться не более чем двумя признаками: морфологическими тегами и тегами лемм, или лексико-семантическими тегами и тегами лемм. Например,

(7) S f inan pl acc|*слеза* + УМИЛЕНИЕ
t:stuff r:concr t:liq|*слеза* + УМИЛЕНИЕ

(8) t:word r:concr r:abstr|*слово* + БЛАГОДАРНОСТЬ
S n inan pl ins|*слово* + БЛАГОДАРНОСТЬ
S n inan pl acc|*слово* + БЛАГОДАРНОСТЬ

(9) A,norm=acc,sg,f,plen|*опасный, тупой, механический, средневековый* + БРИТВА
A,norm=(gen,sg,f,plen|dat,sg,f,plen|ins,sg,f,plen|loc,sg,f,plen)|*опасный, безопасный* + БРИТВА
r:rel ev|*опасный* + БРИТВА
r:rel ev d:neg der:a|*безопасный* + БРИТВА
r:rel der:s|*механический* + БРИТВА

Одна из особенностей формата выдачи данных о конструкциях связана с тем, что у служебных слов (FW) отсутствует семантическая разметка, поэтому среди конструкций регулярно встречаются структуры вида FW+*lex*+FW, хуже интерпретируемые в комбинациях с лексико-семантическими тегами, но более очевидные в комбинациях с морфологическими тегами. Например, конструкции FW + ПОХВАЛА + FW может соответствовать структура типа PR|*к, за, после, в, с* + ПОХВАЛА + PR|*против, сверх, сквозь, в, на*.

Заметим, что большой интерес вызывают конструкции с компонентами, в состав которых входят лексико-семантические теги, поскольку чаще всего с ними ассоциируются группы лемм, выражающих общее значение и характеризующихся близкими дистрибутивными свойствами. Например:

(10) r:rel|*риторический, мировой, процедурный, спорный, шекспировский, практический, методический* + ВОПРОС
ОБСУЖДЕНИЕ + t:ment r:abstr|*проект, концепция* + r:abstr|*благоустройство, реформирование, реформа*
ОТВЕТ + FW + t:speech r:abstr|*запрос, призыв, вопрос, приветствие, просьба, высказывание, похвала, рапорт, реплика*

Наши данные позволяют проследить развертку простейшей структуры в сложную многокомпонентную конструкцию и исследовать видоизменение состава конструкции по пути движения от простого к сложному. Например,

(11) t:poss|*дать, получить, давать* + ОТВЕТ
r:qual|*простой, неточный, точный, вероятный, логичный, нужный, вразумительный, ясный, приличный* + ОТВЕТ
r:rel|*готовый, однозначный, стандартный, истинный, числовой, заданный, релевантный, эмоциональный, содержательный, необязывающий, отрицательный, утвердительный, хлесткий, окончательный, известный, конкретный, официальный, адекватный, отечественный, обстоятельный, определенный, реактивный, обоснованный, очевидный, зачаточный, энергичный, соответствующий, стойкий* + ОТВЕТ
t:move t:poss|*найти* + r:qual|*простой, точный, приличный* + ОТВЕТ + FW + t:speech r:abstr|*вопрос*
t:poss|*давать, дать* + r:rel|*конкретный, однозначный, окончательный* + ОТВЕТ + FW + *вопрос*

ОТВЕТ + PR|*на* + t:speech r:abstr|*приветствие, вопрос, высказывание, рапорт, реплика*
V pf tran inf act|*найти, дать* + A m sg acc inan plen|*простой, однозначный* + ОТВЕТ + PR|*на* ++S m inan sg acc|*вопрос*
найти, дать + простой, однозначный + ОТВЕТ + *на* + *вопрос*

Программа выделения конструкций позволяет получить основные статистические данные об их встречаемости в корпусе (абсолютные и относительные частоты), например:

(12) лемма: РЕКОМЕНДАЦИЯ
объем выборки: 193 контекста

sem+lex+sem
22 (19,13 %) FW + РЕКОМЕНДАЦИЯ + FW
13 (11,30 %) r:rel|*методический, негласный, технологический, подробный, конкретный, методологический* + РЕКОМЕНДАЦИЯ + FW
8 (6,95 %) r:poss|*свой, их, его, наш, мой, ее* + РЕКОМЕНДАЦИЯ + FW
4 (3,47 %) FW + РЕКОМЕНДАЦИЯ + t:hum r:concr t:prof|*специалист, политолог, стоматолог, журналист*
4 (3,47 %) FW + РЕКОМЕНДАЦИЯ + t:hum r:concr|*старик, лекарь, спортсмен, член*
3 (2,60 %) t:poss|*получить, давать, дать* + РЕКОМЕНДАЦИЯ + FW
2 (1,73 %) r:dem|*этот* + РЕКОМЕНДАЦИЯ + r:rel|*особый, национальный*
2 (1,73 %) t:speech r:abstr|*просьба, совет* + РЕКОМЕНДАЦИЯ + FW
2 (1,73 %) r:abstr|*поступление, написание* + РЕКОМЕНДАЦИЯ + FW
2 (1,73 %) r:abstr t:be:appear|*разработка* + РЕКОМЕНДАЦИЯ + FW

gr+lex+gr
6 (5,21 %) CONJ|*и, однако* + РЕКОМЕНДАЦИЯ + PR|*по, в*
4 (3.47 %) A pl gen plen|*общий, подробный, конкретный, методический* + РЕКОМЕНДАЦИЯ + PR|*на, по*
4 (3,47 %) PR|*по* + РЕКОМЕНДАЦИЯ + APRO m sg gen|*этот, ваш, свой, один*
3 (2,60 %) A pl nom plen|*методический, негласный* + РЕКОМЕНДАЦИЯ + PR|*по, на*
3 (2,60 %) PR|*к, согласно, по* + РЕКОМЕНДАЦИЯ + S m anim sg gen|*политолог, производитель, журналист*
3 (2,60 %) A pl ins plen|*методический, полезный* + РЕКОМЕНДАЦИЯ + PR|*по, относительно*
3 (2,60 %) APRO pl acc inan|*свой, весь* + РЕКОМЕНДАЦИЯ + PR|*по, о*
3 (2,60 %) PR|*на, за* + РЕКОМЕНДАЦИЯ + PR|*по, к*
2 (1,73 %) V pf tran inf act|*подготовить, разработать* + РЕКОМЕНДАЦИЯ +PR|*о, по*
2 (1,73 %) A pl acc inan plen|*соответствующий, лестный* + РЕКОМЕНДАЦИЯ + S m anim sg dat|*оператор, кот*

5. Визуализация структуры и наполнения конструкций

Наша нынешняя задача — из многообразия используемых в компьютерной лингвистике техник визуализации (ср., например, [Penn et al. 2009]) выбрать метод графического представления данных, отличающийся простотой и широкими иллюстративными возможностями, позволяющий отразить как состав конструкций, так и иерархию их компонентов.

Для получения графических представлений, отражающих структуру и наполнение конструкций, был задействован модуль pattern.graph (http://www.clips.ua.ac.be/pages/pattern-graph, [De Smedt 2012]), разработанный на языке Python и предназначенный для визуализации различных типов связей в тексте на естественном языке. На входе он принимает строку, обозначающую конструкцию, в формате, описанном в Разделе 2. На выходе создается граф, иллюстрирующий соответствующую конструкцию (см. рис. 2–3).

Визуализация производится в два этапа: во-первых, производится парсинг строки конструкции и выявление ее главных и второстепенных элементов с сохранением порядка следования, причем главными являются элементы, которые необходимым образом присутствуют и в своей линейной последовательности формируют конструкцию, в то время как второстепенные представляют из себя парадигматические варианты наполнения главных; формат входной строки позволяет однозначно провести данное разграничение; во-вторых, из них создается граф, отражающий данные структурные соответствия между элементами.

5.1. Парсинг конструкции и выявление ее элементов

В качестве исходных данных служат строки такого вида, как в примерах (1) и (4). В строках выявляются главные и второстепенные элементы. В главные элементы входит, во-первых, целевое слово, не имеющее тегов разметки; а также грамматические или лексико-семантические теги остальных слов в конструкции. Соответственно, первостепенными элементами в примерах (1) и (4) будут следующие (с учетом порядка следования):

(13) ОТВЕТ; PR; t:speech r:abstr

(14) ЛЮБОВЬ; PR; ANUM m sg gen; S f inan sg gen

Второстепенными элементами считаются леммы, обозначающие наполнение первостепенных грамматических и лексико-семантических элементов. Их количество может варьироваться от одного до нескольких десятков. Ср. соответствующие второстепенные элементы в примерах (1) и (4):

(15) *на*; *приветствие, вопрос, высказывание, рапорт, реплика*

(16) *с*; *первый; взгляд*

Для каждого второстепенного элемента сохраняется его соответствие «родному» первостепенному тегу.

Полученная структура из первостепенных, второстепенных элементов, порядка следования для первых и связей между первыми и вторыми передается для визуализации.

5.2. Визуализация элементов конструкции

В качестве узлов графа изображаются первостепенные и второстепенные элементы. При этом закрашенными узлами обозначаются главное слово конструкции, которое дополнительно выделяется красным цветом; а также лексемы, отражающие лексическое наполнение конструкции. Узлы лексико-семантических и морфологических тегов остаются пустыми.

Первостепенные элементы связываются направленными ребрами графа, в соответствии с порядком следования этих элементов в конструкции. Второстепенные элементы связываются с тегами, которые они наполняют, с помощью двунаправленных ребер. Для наглядности стрелки первого типа подсвечиваются зеленым, второго — синим цветом.

Задействованный нами модуль pattern.graph обладает особым удобством в использовании, так как не требует вычисления координат для изображения узлов и ребер. При указании размера узлов, наличия, длины и толщины ребер, их местонахождение вычисляется автоматически.

Примеры визуализации данных о конструкциях средствами модуля pattern.graph приведены на рис. 2–3.

Структура конструкции в графах отражается следующим образом: красным цветом помечен узел, содержащий целевое слово, зеленым цветом выделены ребра графа, связывающие между собой элементы разметки конструкции (лексико-семантические и морфологические теги), синим — ребра графа, связывающие теги лемм с лексико-семантическими и морфологическими тегами.

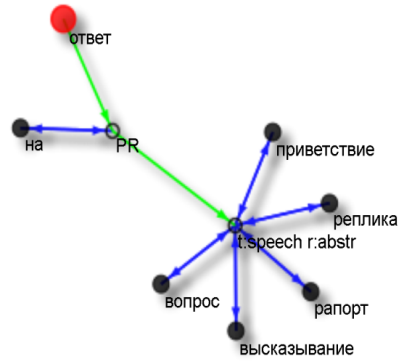

Рис. 2. Графическое представление конструкции ОТВЕТ + PR|*на* + t:speech r:abstr| *приветствие, вопрос, высказывание, рапорт, реплика*

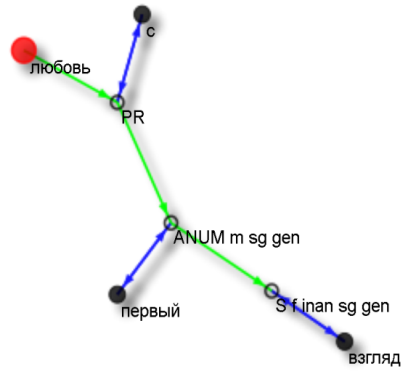

Рис. 3. Графическое представление конструкции
любовь + PR|*с* + ANUM m sg gen|*первый* + S f inan sg gen|*взгляд*

6. Заключение

Проведенные эксперименты дают основания утверждать, что

1) инструмент автоматического выделения конструкций приспособлен для обработки контекстных выборок из НКРЯ, его применение позволило получить списки конструкций для целевых существительных из лексико-семантических групп названий инструментов, обозначений речевых действий и названий эмоций;
2) полученные конструкции различаются по числу компонентов (это пары, тройки, четверки, пятерки, состоящие из тегов контекстного окружения) и по наполнению (это двухслойные структуры, в состав которых входят либо морфологические теги и теги лемм, либо лексико-семантические теги и теги лемм);
3) задача визуализации данных о выделенных конструкциях успешно решается с помощью модуля pattern.graph, позволяющего наглядно представлять организацию конструкций, иерархию и различные типы их компонентов.

В перспективе, мы бы хотели рассмотреть возможность представлять в конструкции три слоя разметки (леммы, грамматические теги, лексико-семантические теги) одновременно. Кроме того, хотелось бы учитывать статус факультативных элементов конструкции — в нынешней версии такой функционал не предусмотрен.

Модуль визуализации будет совершенствоваться с учетом пожеланий пользователей. Одной из дальнейших задач видится переход к динамической организации модуля визуализации — особенно в тех случаях, когда конструкции содержат много элементов и много лексических вариантов реализации. Предполагается рассмотреть вопрос о визуальном представлении нескольких конструкций в контексте, когда конструкции с разными лексическими

центрами «наслаиваются» друг на друга. Наконец, планируется провести сопоставление выделенных наборов лексических конструкций с наборами, который мог бы выделить лексикограф на тех же данных.

Литература

1. *Fillmore Ch. J., Kay P., O'Connor M. C.* (1988a), Regularity and idiomaticity in grammatical constructions: The case of "let alone", Language, Vol. 64–3.
2. *Fillmore Ch. J.* (1988b), The Mechanisms of Construction Grammar, Proceedings of the Berkeley Linguistic Society, Vol. 14.
3. *Firth J. R.* (1957/1968), A synopsis of linguistic theory 1930–1955, in Palmer F. R. (ed.), Selected Papers of J. R. Firth 1952–1959, Longman, London.
4. *Goldberg A. E.* (1995), Constructions. A Construction Grammar Approach to Argument Structure. University of Chicago Press, Chicago, IL/London.
5. *Goldberg A. E.* (2006), Constructions at Work: the Nature of Generalization in Language, Oxford University Press, Oxford.
6. *Gries St. Th., Divjak D. S.* (2009), Behavioral profiles: a corpus-based approach towards cognitive semantic analysis, in Evans V., Pourcel S. S. (eds.), New directions in cognitive linguistics, John Benjamins, Amsterdam & Philadelphia, pp. 57–75.
7. *Lyashevskaya O. A., Mitrofanova O. A., Grachkova M. A., Shimorina A. S., Shurygina A. S., Romanov S. V.* (2012), Building the Inventory of Russian nominal Constructions [K postrojeniju inventar'a russkih imennyh konstrukcij], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2012» [Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog 2012»], RGGU, Moscow.
8. *Mitrofanova O. A., Lyashevskaya O. A., Grachkova M. A., Shimorina A. S., Shurygina A. S., Romanov S. V.* (2012), Experiments on Automatic Word Sense Disambiguation and Construction Identification (Based on Russian National Corpus) [Eksperimenty po avtomaticheskomu razresheniju leksiko-semanticheskoj neodnoznachnosti i vydeleniju konstrukcij (na materiale Nacional'nogo korpusa russkogo jazyka)], Strukturnaja i prikladnaja lingvistika. Vyp. 9 [Structural and Applied Linguistics. Vol. 9], St. Petersburg.
9. *Penn G., Carpendale Sh., Collins Chr.* (2009), Interactive Visualization for Computational Linguistics: Tutorial at ESSLLI-09, available at: esslli2009.labri.fr/documents/carpendale_penn.pdf.
10. *Sahlgren M.* (2008), The Distributional Hypothesis, Rivista di Linguistica [Italian Journal of Linguistics], Vol. 20 (1), pp. 33–53.
11. *de Smedt T., Daelemans W.* (2012), Pattern for Python, Journal of Machine Learning Research, Vol. 13.
12. *Tomasello M.* (2003), Constructing a Language: A Usage-Based Approach to Child Language Acquisition, Harvard University Press, Cambridge, MA.