К ПОСТРОЕНИЮ ИНВЕНТАРЯ РУССКИХ ИМЕННЫХ КОНСТРУКЦИЙ[1]

Ляшевская О. Н. (olesar@gmail.com)
Национальный исследовательский университет
Высшая школа экономики, Москва, Россия

Митрофанова О. А. (alkonost-om@yandex.ru)
Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

Грачкова М. А. (maaag86@mail.ru)
Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

Романов С. В. (complefor@rambler.ru)
ЗАО «Интернет-Проекты», Санкт-Петербург, Россия

Шиморина А. С. (shinas@yandex.ru)
Институт лингвистических исследований РАН,
Санкт-Петербург, Россия

Шурыгина А. С. (sanyana@gmail.com)
Российский государственный гуманитарный университет,
Москва, Россия

В докладе обсуждаются результаты экспериментов по автоматическому выделению конструкций, проводимых на материале Национального корпуса русского языка (НКРЯ). С этой целью разработан компьютерный инструмент, позволяющий извлекать и обрабатывать сочетаемостные данные из выборок НКРЯ. В качестве целевых слов выступают русские имена существительные. Для каждого из целевых слов получены списки конструкций — наиболее частотные сочетания, включающие целевое слово, частотные лексико-семантические теги — контекстные маркеры того или иного значения целевого слова, а также частотные леммы, характеризуемые этими семантическими тегами. Например: ВИД (*разновидность, тип*) + **r:abstr t:sport:** *спорт, футбол, биатлон* и т. д. Выделенные конструкции систематизируются по структуре и лексико-семантическому наполнению. В заключение проводится проверка результатов экспериментов, предполагающая сравнение списков конструкций со сведениями о коллокациях, устойчивых сочетаниях и т. д., зарегистрированных в различных лингвистических источниках (сервисы поиска биграмм, словари).

Ключевые слова: конструкции, корпус текстов, лексико-семантическая разметка, существительные, русский язык

---

[1] Работа выполнена при финансовой поддержке РФФИ (проект 10-06-00586-а) и программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» (проект FrameBank), а также проекта НИР «Модель интегрированного программно-лингвистического комплекса для создания специализированных корпусов русского языка».

BUILDING THE INVENTORY OF RUSSIAN NOMINAL CONSTRUCTIONS

Lashevskaya O. N. (olesar@gmail.com),
National Research Institute Higher School of Economics, Moscow, Russia

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

Grachkova M. A. (maaag86@mail.ru),
Saint-Petersburg State University, Saint-Petersburg, Russia

Romanov S. V. (complefor@rambler.ru),
ZAO «Internet-Projects», Saint-Petersburg, Russia

Shimorina A. S. (shinas@yandex.ru),
Institute of Linguistic Studies RAS, Saint-Petersburg, Russia

Shurygina A. S. (sanyana@gmail.com), Russian State University for Humanities, Moscow, Russia

The paper presents experimental results on automatic construction identification performed on the Russian National Corpus (RNC). For this purpose we developed a toolbox which allows to extract and process co-occurrence data from RNC samples. Russian nouns are chosen as target words. Lists of constructions were built for each target word. By constructions we mean frequent word combinations which include a target word and frequent lexical-semantic tags — context marker of certain meanings of a target word, as well as frequent lemmas representing the given lexical-semantic tags. E.g.: ВИД (*kind, sort, type*) + **r:abstr t:sport:** *спорт (sport), футбол (football), биатлон (biathlon),* etc. Extracted constructions are grouped according to their structure and lexical-semantic content. In conclusion we perform verification of experimental results which implies comparison of lists of constructions with lists of collocations, idioms, etc. registered in various linguistic resources (bigram search engines, dictionaries).

Key words: constructions, text corpus, lexical-semantic annotation, nouns, Russian

1. Введение

Создание лексикографических ресурсов на основе корпусов текстов — одно из продуктивных направлений современных исследований в области автоматической лексикографии и компьютерной лингвистики. Главное достоинство корпуса текстов как источника лингвистических данных заключается в том, что на его основе возможно создание не только традиционных ресурсов (в частности, словарных баз данных), но и ресурсов, объектом описания в которых являются довольно сложные лингвистические явления, в частности, конструкции (см., например, **http://dict.ruslang.ru/**, Kustova 2011). Так, автоматическое построение каталога русских конструкций на базе Национального корпуса русского языка (НКРЯ, **http://www.ruscorpora.ru/**) — цель нашего проекта, реализуемого коллективами НКРЯ и кафедры математической лингвистики СПбГУ.

В исследованиях, опирающихся на конструкционные подходы в лингвистике, используется широкое понимание конструкции: в общем случае, это сложный знак, значение которого не выводится из значения составляющих и компоненты которого взаимодействуют между собой и взаимообуславливают друг друга (Rakhilina 2010; Oskol'skaya, Say 2010, Ovsjannikova et al. 2011). Лексически-ориентированные конструкции предполагают организацию вокруг одного или нескольких фиксированных лексических элементов (например, коллокация, коллострукция, морфо-синтаксическая схема, модель управления, комбинация с лексической функцией, синтаксическая группа с фиксированными лексическими слотами, или конструкция «малого синтаксиса», и др.).

Идентификация лексических конструкций в корпусе связана с двумя важными задачами. Во-первых, сочетаемость и синтаксис довольно больших слоев лексики (прежде всего, имен существительных, прилагательных и наречий) описана в русистике совершенно недостаточно (из положительных примеров можно назвать TKS 1984, Apresjan 2003, 2010 и нек. др., охватывающие лишь отдельные единицы). Во-вторых, процедура выделения конструкций связана с разрешением лексико-семантической неоднозначности (Mitrofanova et al. 2008; Mitrofanova, Lyashevskaya 2009; Lyashevskaya et al. 2011; Shimorina, Grachkova 2011). Если слово многозначно, то, в идеале, каждое значение характеризуется своим кругом конструкций — тем самым, анализ семейств близких конструкций дает информацию, позволяющую распознавать и разграничивать значения многозначного слова. Кроме того, эксперимент с автоматическим выделением конструкций позволяет проверить теоретическую гипотезу о генерализациях в ходе усвоения языка ребенком (Dąbrowska 2004). Эта гипотеза гласит, что человеку свойственно собобщать сходные часто повторяемые цепочки слов в виде единиц более абстрактного уровня (=конструкций), причем ключами для генерализации могут служить единицы морфологического, лексемного, синтаксического и любого другого уровня.

Таким образом, под конструкцией в нашем проекте понимается сочетание целевого слова и контекстных маркеров его значения, характеризующееся частотностью и устойчивостью. Данная трактовка конструкций согласуется с основными идеями Грамматики конструкций (Fillmore 1988; Goldberg 1995,

2006; Tomasello 2003; Kuznetsova 2007). Как контекстные маркеры рассматриваются теги, доступные в многоуровневой разметке контекстов НКРЯ: теги лемм (*lex* — лексема, которой принадлежит словоформа), морфологические теги (*gr* — грамматические признаки словоформ: частеречная принадлежность, значения грамматических категорий и т. д.), лексико-семантические теги (*sem* — признаки, указывающие на принадлежность слова к определенному лексико-семантическому классу[2]).

Примерами конструкций могут служить следующие сочетания существительного *вид* с левосторонними/правосторонними соседями, маркирующими различные его значения:

(1) значение **r:abstr t:perc der:v**: *внешность, видимый облик; состояние внешний* + ВИД

ВИД + **r:concr t:hum:** *человек, женщина, король, друг, сосед, господин, товарищ* и т. д.

(2) значение **r:abstr r:concr pt:set sc:X**: *разновидность, тип*

r:card:pauc: *два, три, четыре, оба* и т. д. + ВИД

r:abstr der:v: *использование, стоимость, смешение, доминирование, сохранение* и т. д. + ВИД

ВИД + **r:abstr t:sport:** *спорт, футбол, физкультура, картинг, биатлон* и т. д. и др. В качестве более сложных конструкций могут выступать такие цепочки, как *представить* + Sacc + *В* + *ВИДЕ* + Sgen.

В настоящее время ведется активное обсуждение методов и алгоритмов выделения конструкций (Sahlgren, Knutsson 2009; Proceedings of the NAACL… 2010; Wible, Tsao 2010). Большие успехи достигнуты в области извлечения *n*-грамм (коллокаций, неоднословных целостностей — ср. (Manning, Schütze 2002, среди многих других; для русского языка –Jagunova, Pivovarova 2011). Однако идиоматизированные конструкции, а также конструкции с нестандартной синтаксической структурой, хотя они подробно описаны в исследовательской литературе (Borisova 1995, Iordanskaya, Mel'chuk 2007), представляют серьезную проблему в автоматической обработке текста. Автоматизация выделения конструкций из русскоязычных текстов усложняется такими особенностями материала, как свободный порядок слов и богатое словоизменение, снижающие предсказуемость формальной организации конструкций. По этой причине выделение конструкций с опорой на *n*-граммы, допустимое, например, в англоязычных текстах, в нашем случае не всегда приводит к желаемым результатам:

---

[2] Например, **r:concr** — предметные имена (*девочка, стол, молоко*); **r:abstr** — непредметные имена (*вождение, яркость, время*); **t:hum** — лица (*человек, учитель*); **t:stuff** — вещества и материалы (*вода, песок, тесто, жесть, шелк*); **t:constr —** здания и сооружения (*дом, шалаш, мост*); **t:tool** — инструменты и приспособления (*молоток, палка, пуговица, машина*); **t:move** — движение (*беготня, вынос, качка*); **t:perc** — восприятие (*осязание, слух, видимость, взгляд, зрелище*); **t:space** — пространство и место (*космос, город, тайга, овраг, вход*); **t:psych** — психическая сфера (*апатия, безумие, вдохновение, спокойствие*); **t:speech** — речь (*дискуссия, молва, ахинея, реплика, подковырка*); **r:qual** — качественные (*хороший, большой*); **r:rel** — относительные (*деревянный, лунный*); и т. д., подробнее см.: **http://www.ruscorpora.ru/corpora-sem.html.**

фрагменты конструкций могут не иметь явного выражения в рамках заданного контекста, могут выходить за пределы синтаксических групп и т. д.

Существует ряд проектов, в которых особое внимание уделяется формализации лексико-синтаксических связей единиц текста, среди них есть исследования и на материале русского языка: Word Sketches для русского языка (Zakharov, Khohlova 2010), работы по извлечению лексико-синтаксических шаблонов (Bolshakova et al. 2007), по автоматическому построению словарей сочетаемости (Gel'buh et al. 2004). Тем не менее, электронные ресурсы, отражающие сочетаемостные предпочтения и рамки валентностей русской лексики, недостаточно разработаны. Если бы в распоряжении исследователей были такие ресурсы, как FrameNet или PropBank для русского языка, то это существенно облегчило задачу каталогизации конструкций (Lyashevskaya, Kuznetsova 2009).

В данной статье обсуждается один из важнейших аспектов проекта, а именно, построение шаблонов именных конструкций на основе контекстных маркеров значений целевых слов.

2. Лингвистические данные

Источником данных об именных конструкциях служат выборки контекстов из НКРЯ, крупнейшего русскоязычного корпуса с многоярусной разметкой. В центре внимания исследовательской группы находятся (1) отдельные русские существительные, представляющие различные лексико-семантические группы: *дом*, *вид*, *орган, лук, глава* и т. д., (2) целостные лексико-семантические группы существительных — названий инструментов (*бритва, веник, весло, карандаш, коса, лом, лопата, метла, ножницы, топор, щетка*), обозначений речевых действий (*дискуссия, комплимент, обращение, обсуждение, ответ, похвала, рекомендация, вопрос, вранье, выражение*) и т. д. Анализируемые лексемы отличаются количеством значений, характером развития полисемии/омонимии, степенью связанности значений между собой. В нашем исследовании производится условное приравнивание омонимичных коррелятов к многозначным словам (Rakhilina et al. 2006). Разметка значений слов в контекстах НКРЯ проводилась на основе Семантического словаря НКРЯ. В табл. 1 приводится филиация значений слова *глава* с указанием их лексико-семантической аннотации и частотности в выборке, в табл. 2 — пример разметки контекстов со словом *глава*.

Таблица 1. Филиация значений слов *глава*

Значения	Лексико-семантическая аннотация	Примеры	Число контекстов: 1056, из них
т1. То же, что голова	**r:concr pt:partb pc:hum** предметное имя, часть тела человека	*Склонить главу.*	8
т2. ***Во главе*** *— возглавляя кого-что-н., впереди кого-чего-н.* ***Во главу угла ставить*** *— считать самым важным*	**r:concr der:shift dt:partb** предметное имя, метонимический перенос	*Идти во главе колонны.*	136
т3. Купол церкви	**r:concr pt:part pc:constr** предметное имя, часть здания	*Главы собора.*	1
т4. Руководитель, начальник, старший по положению	**r:concr t:hum** предметное имя, лицо	*Глава государства.*	299
т5. Раздел книги, статьи	**r:concr t:text pt:part pc:text** предметное имя, текст, часть текста	*Глава книги.*	612

Таблица 2. Пример разметки контекстов со словом *глава*

Значение	*т5. Раздел книги, статьи*
Контекст	*Глава 8, в которой родители получают странные письма.*
Разметка	<w><ana lex='глава' gr='S,f,inan=sg,nom' SEMF='r:concr t:text pt:part pc:text '/></ana>Глав`а</w> <w><ana lex='8' gr='NUM=ciph'/></ana>8</w>, <w><ana lex='в' gr='PR'/></ana>в</w> <w><ana lex='который' gr='A-PRO=f,sg,loc' sem='r:rel | r:rel t:ord '/></ana>кот`орой</w> <w><ana lex='родитель' gr='S,m,anim=pl,nom' sem='r:concr t:hum:kin d:nag der:v '/></ana>род`ители</w> <w><ana lex='получать' gr='V,ipf,tran=pl,act,praes,3p,indic,act' sem='d:pref der:v '/></ana>получ`ают</w> <w><ana lex='странный' gr='A=pl,acc,plen' sem='r:qual '/></ana>стр`анные</w> <w><ana lex='письмо' gr='S,n,inan=pl,acc' sem='r:concr t:text'/></ana>п`исьма</w>

3. Компьютерный инструмент выделения конструкций

Компьютерный инструмент выделения конструкций позволяет выполнять автоматическую классификацию контекстов, с этой целью создается векторная модель экспериментальной выборки; в качестве базового алгоритма выбран алгоритм классификации с учителем. Данные процедуры осуществляются с помощью программного обеспечения, разрабатываемого С. В. Романовым на языке Python. Инструмент работает в двух режимах: формирование классов контекстов, соотносимых с отдельными значениями целевого слова; генерация списков кандидатов в конструкции — наиболее частотных сочетаний, в которых реализуется то или иное значение целевого слова. Автоматическое выделение конструкций производится на основе статистических данных о сочетаемости целевых слов и левосторонних/правосторонних соседей — контекстных маркеров их значений: тегов *lex*, *gr* и *sem*. Сочетаемостная информация извлекается из обучающей выборки. Возможно варьирование таких параметров экспериментов, как ширина контекстного окна [*-l*; *+r*], обработка с учетом/без учета весов контекстных элементов. Результат работы программы отражается в виде списка частотных комбинаций целевого слова и статистически значимых контекстных маркеров, с указанием данных о частоте встречаемости этих сочетаний и с перечнями лексем, реализующих значения контекстных маркеров в их составе. Пример выдачи программы приведен в табл. 3. Подобные списки подлежат постредактированию, в ходе которого отсеиваются сочетания, малоинформативные с точки зрения определения значения целевого слова в контексте (например, для целевого слова *ножницы* таковыми признаны сочетания с местоимениями).

Таблица 3. Пример выдачи списка кандидатов в конструкции для слова ***ножницы***

Параметры обработки данных	Лемма: *ножницы* Значение: **r:concr t:tool:instr der:s** (*режущий инструмент из двух раздвигающихся лезвий с кольцеобразными ручками*) Контекстное окно: [-1; +1] Объем выборки: 683 контекстов
Кандидаты в конструкции с левосторонними соседями	**t:impact ca:caus d:root:** резать(10) кромсать(1) стричь(4) **t:poss ca:caus:** взять(17) **r:dem:** этот(11) такой(1) тот(1) **t:impact ca:caus d:pref der:v:** вырезать(9) обрезать(4) надрезать(1) обстригать(1) **r:rel der:s dt:space:** садовый(19) …
Кандидаты в конструкции с правосторонними соседями	**r:poss:** ее(3) свой(1) наш(2) мой(2) **r:rel | r:rel t:ord:** который(8) **r:spec:** каждый(1) сам(1) весь(5) **t:impact ca:caus d:pref:** нарезать(2) срезать(1) перерезать(3) искромсать(1) **t:impact ca:caus d:pref der:v:** вырезать(4) надрезать(3) обрезать(1) **r:rel:** что(3) известный(1) дикий(1) лишний(1) обыкновенный(1) **r:pers:** она(2) я(3) он(1) они(2) мы(1) вы(1) …

4. Предварительные результаты экспериментов по выделению конструкций

Эксперименты по выделению конструкций проводятся в несколько этапов. Сначала для каждого значения рассматриваемых целевых слов составляется список контекстов его употребления, далее из контекстов автоматически извлекается наиболее частотная лексико-семантическая и морфологическая информация о контекстных маркерах значения в заданном окне. Выделение конструкций изначально производится в пределах узкого контекстного окна [-1; +1], где высока вероятность встретить контактные контекстные элементы, входящие в устойчивые словосочетания с исследуемыми словами. Рассматриваются также и другие параметры контекстного окна, важные для диагностики дистантных контекстных маркеров и выявления конструкций со структурой, более сложной по сравнению с биграммами, в связи с этим в проводимых экспериментах окно расширяется до [-3;+4]. Далее формируются морфологические модели конструкций, описывается

их лексико-семантическое наполнение. Методика лингвистического анализа морфологических моделей конструкций и их лексико-семантического наполнения, использованная в нашем исследовании, основывается на опыте анализа сочетаемостных предпочтений слов разных частей речи (Mitrofanova, Belik, Kadina 2008).

Очевидно, что отношения между элементами конструкций не сводимы к аддитивным. В ряде случаев конструкции реализуют разнообразные лексические функции (Mel'chuk 1999/1974, 2007), как, например, конструкции, выделенные для слова *комплимент* из лексико-семантической группы речи (значение **t:speech r:abstr**: *любезные, приятные слова, лестный отзыв*):

Mult (название типовой «частичной» совокупности): **pt:set r:concr:** *куча* + комплимент

Ver (правильный, соответствующий, какой следует): **ev r:rel:** *уместный* + комплимент

AntiVer **r:rel:** *сомнительный* + комплимент

$Oper_1$ (глагол, связывающий название первого актанта в роли подлежащего с названием ситуации в роли первого дополнения): **d:pref der:v:** *выражать* + комплимент; **d:root:** *делать* + комплимент.

Такие конструкции были выявлены в выборках из НКРЯ и зафиксированы в ТКС (TKS 1984).

При анализе контекстных данных, полученных для инструментальных существительных, мы наблюдаем разнообразие конструкций с точки зрения сложности (двучленные, трехчленные и т. д.), степени устойчивости, морфосинтаксических свойств. Выявленные конструкции допускают обобщение до шаблонов, реализующих инструментальное значение.

Чаще всего встречаются двучленные конструкции — свободные и устойчивые сочетания вида *A+S / S+A, V+S / S+V, S+S* с различным лексико-семантическим наполнением.

(1) *A+S / S+A*: в таких сочетаниях выражаются
- разнообразные свойства инструмента: **r:rel ev d:neg der:a:** *безопасный* + БРИТВА; **r:rel ev:** *опасный* + БРИТВА; **r:qual t:physq ev:** *острый* + БРИТВА; **r:qual t:physst:** *сухой, мокрый* + ВЕНИК; **r:qual t:physq:** *мягкий, жесткий, гибкий, твердый* + ЩЕТКА; **r:qual t:physq:color der:s dt:color dt:abstr:** *цветной* + КАРАНДАШ и т. д.;
- тип инструмента: **r:rel der:s:** *механический, электрический* + БРИТВА; **r:rel der:s:** *химический* + КАРАНДАШ; **r:rel der:s dt:tool:instr dt:tool:** *совковый* + ЛОПАТА и т. д.;
- субъекты, использующие инструменты: **r:rel der:s dt:hum:** *саперный, дворницкий* + ЛОПАТА и т. д.;
- сфера применения инструмента: **r:rel der:s dt:space:** *садовый* + НОЖНИЦЫ; **r:rel der:s:** *маникюрный, хирургический* + НОЖНИЦЫ; **r:rel der:s dt:tool:cloth dt:tool:** *платяной, сапожный, обувной* + ЩЕТКА и т. д.;
- материал, из которого изготовлен инструмент: **r:rel der:s dt:stuff:** *деревянный, фанерный, железный, пластмассовый, металлический* +

ЛОПАТА; **r:rel der:s dt:stuff:** *металлический, пластиковый, железный, пластмассовый, капроновый* + ЩЕТКА; *S+A*: ЩЕТКА + **r:rel der:s dt:stuff:** *пенный, абразивный* и т. д.;

– размер инструмента: **r:qual t:size:max:** *широкий, огромный, большой* + ЛОПАТА и т. д.;

– форма инструмента: **r:qual t:physq:form:** *круглый, плоский* + ЩЕТКА и т. д.

(2) *V+S / S+V*: в таких сочетаниях выражаются

– действия, производимые при помощи инструмента: БРИТВА + **t:impact ca:caus d:pref:** *срезать*; БРИТВА+ **t:impact ca:caus d:pref | t:impact ca:caus d:pref**: *разрезать*; ВЕНИК + **d:pref der:v | t:move ca:caus d:pref der:v:** *подметать*; **t:move d:pref der:v:** *огребать, подгребать* + ВЕСЛО; **d:impf d:pref der:v:** *исписывать, подчеркивать* + КАРАНДАШ; **d:pref der:v:** *нарисовать, отмечать* + КАРАНДАШ; КАРАНДАШ + **d:root:** *писать*; КАРАНДАШ + **d:pref der:v:** *помечать, нарисовать, изображать*; КАРАНДАШ + **d:impf d:pref der:v:** *подчеркивать, записывать*; ЛОПАТА + **d:impf d:pref der:v:** *откапывать, разравнивать, расковыривать, раскидывать, перемешивать, перекапывать*; **t:move ca:caus d:root:** *махать, гонять, мести* + МЕТЛА; МЕТЛА + **t:move ca:caus d:root:** *мести, гонять*; МЕТЛА + **t:move d:pref der:v:** *сметать, выгонять, сгребать, отгонять*; **t:impact ca:caus d:root**: *резать, стричь, кромсать, ранить* + НОЖНИЦЫ; **t:impact ca:caus d:pref der:v:** *вырезать, обрезать, обстригать, надрезать* + НОЖНИЦЫ; НОЖНИЦЫ + **t:impact ca:caus d:pref:** *перерезать, нарезать, срезать, взрезать, откромсать, искромсать, сломать, прорезать*; НОЖНИЦЫ + **t:impact ca:caus d:pref der:v:** *обрезать, вырезать, надрезать*; НОЖНИЦЫ +**t:impact ca:caus d:root:** *кромсать, резать, стричь*; **t:poss ca:caus:** *взять* + НОЖНИЦЫ; **t:impact d:pref der:v:** *вырубать, отрубать, приглушать, разрубать, обрубать, врубать, изрубать* + ТОПОР; ТОПОР + **t:impact ca:caus d:root:** *бить, рубить, тесать*; ТОПОР + **t:impact d:pref der:v:** *разрубать, отрубать, обрубать, скреплять, прорубать, вырубать*; **t:poss ca:caus:** *взять* + ТОПОР и т. д.

– действия, производимые применительно к инструменту: ВЕНИК + **d:impf d:pref der:v:** *ошпаривать, раскидывать, наготавливать, заготавливать, пересушивать*; весло + **d:root:** *бросать*; МЕТЛА + **t:impact:creat d:root:** *вязать* и т. д.

(3) *S + S*: в таких сочетаниях выражаются

– совместно используемые инструменты: ВЕНИК + **r:concr t:tool:dish top:contain:** *ведро*; МЕТЛА + **r:concr t:tool:instr:** *лопата*, *швабра*, *щетка*, *веник*; ЛОПАТА + **r:concr t:tool:instr der:v:** *метла, грабли, скребок, тяпка*; ЩЕТКА + **r:concr t:tool:instr der:v:** *расческа, шило, чесалка, скребница, скребок, метла, бритва*; и т. д.

– вещества, материалы, задействованные при использовании инструмента: БРИТВА + **r:concr t:stuff:** *одеколон*; КАРАНДАШ + **r:concr t:stuff:** *ватман, бумага, ластик, акварель*; **r:concr t:stuff:** *уголь, бумага, пастель, ластик* + КАРАНДАШ; ЛОПАТА + **r:concr t:stuff:** *глина, навоз, цемент, уголь, грунт, песок, бетон*; ЩЕТКА + **r:concr t:stuff t:tool pt:qtm qc:stuff der:v:** *мыло*; **r:concr t:stuff:** *паста, гуталин, шампунь, песок, алюминий, вакса, вельвет, бархотка, хрусталь, наст, порошок, одеколон, стекло, ковер, сода,* + ЩЕТКА и т. д.
– субъекты, использующие инструмент: **r:concr t:hum der:s:** *сменщик, девушка, дружинник, сибиряк, мужик* + ВЕСЛО; МЕТЛА + **r:concr t:hum t:prof der:s:** *дворник*; **r:concr t:hum t:prof der:s:** *дворник* + МЕТЛА; ТОПОР + **r:concr t:hum:** *вельможа, человек, крестьянин, сэр, друг, дитя* и т. д.
– объекты, частью которых является инструмент: ВЕСЛО + **r:concr t:tool:transp:** *лодка, ушкуй, байдарка, ладья, крейсер*; **r:concr t:tool:transp:** *галера, корабль, лодка* + ВЕСЛО и т. д.
– части тела человека: БРИТВА + **r:concr pt:partb pc:hum:** *горло, рука, лицо, нога*; КАРАНДАШ + **r:concr pt:partb pc:hum:** *рука*; **r:concr pt:partb pc:hum:** *рука* + ЛОПАТА; ТОПОР + **r:concr pt:partb pc:hum:** *рука, лицо, нога*; ЩЕТКА + **r:concr t:hair pt:aggr sc:hair top:rope:** *волос* и т. д.
– части инструментов: **r:concr t:thing pt:residpart pc:X der:v:** *огрызок* + КАРАНДАШ и т. д.

Для большинства существительных — названий инструментов характерны синтаксические группы с фиксированными лексическими элементами, например, **r:rel t:manner:** *как* + БРИТВА, ВЕНИК, ЛОПАТА, ВЕСЛО и т. п. (сравнительный оборот). В данном случае имеется двучленное свободное сочетание структуры *CONJ* + *S*. В выборках зарегистрированы трех- и четырехчленные фраземы: **t:move ca:caus d:root | d:root:** *гнать* + **r:rel ev:neg:** *поганый* + МЕТЛА (*гнать поганой метлой*); **r:concr t:org:** *фирма* + ВЕНИК + *не* + **t:impact:creat d:root:** *вязать* (*фирма веников не вяжет*).

В завершение осуществляется процедура верификации результатов: полученные списки конструкций сопоставляются со списками коллокаций, формируемыми на основе сервиса поиска биграмм С. А. Шарова (**http://corpus.leeds.ac.uk/ruscorpora.html**), с данными словарей МАС и БАС. Обнаружено, что отдельные варианты лексико-семантического наполнения конструкций характеризуются высокой устойчивостью. Среди контекстных маркеров значений целевых слов в составе конструкций присутствуют компоненты коллокаций с высоким показателем Log-Likelihood (LL). Отдельные коллокации нашли отражение в словарях МАС и БАС как устойчивые сочетания (приводятся с пометой ◊ в БАС) и фразеологизмы (приводятся с пометой ◊ МАС и с пометой ~ в БАС) (см. табл. 4).

Таблица 4. Устойчивые сочетания в составе конструкций со словом *карандаш*

КАРАНДАШ ‘ручное орудие для копания, сгребания с рукояткой и широким плоским отточенным концом’ (r:concr t:tool:instr)		
Конструкции	**Сервис поиска биграмм С. А. Шарова**	**Словари**
r:qual t:physq:color der:s dt:color dt:abstr: *цветной* + КАРАНДАШ		***Цветные*** *карандаши.*
r:concr t:thing pt:residpart pc:X der:v: *огрызок* + КАРАНДАШ		***Огрызок*** *карандаша.*
r:rel der:s: химический+КАРАНДАШ	***химический*** LL = 60,34	***Химический*** *карандаш.*
КАРАНДАШ + **d:root:** *писать*		***Писать*** *карандашом.*
КАРАНДАШ + **r:concr pt:partb pc:hum:** *рука*	***рука*** LL = 36,63	

5. Заключение

Исследование показывает, что задача автоматического выделения конструкций, понимаемых как сочетания целевых слов и контекстных маркеров их значений, выражаемых тегами в многоярусной разметке НКРЯ, имеет практическое решение:

(1) разработан инструмент автоматического выделения конструкций, продуктивность которого подтверждена в сериях экспериментов с различными русскими существительными;

(2) разработана и апробирована методика анализа данных о конструкциях, предложены принципы их систематизации по морфологическим моделям, лексико-семантическому наполнению и ряду других критериев;

(3) проведены пилотные эксперименты по выделению конструкций для русских существительных, получены рабочие списки конструкций для инструментальных существительных, существительных со значением речи и некоторых других слов;

(4) проведена верификация экспериментальных данных: полученные в ходе исследования списки конструкций сравнивались со списками коллокаций для целевых слов, а также с перечнями устойчивых сочетаний и фразеологизмов из словарей русского языка.

Тем самым, лингвистические данные и программные решения, полученные в ходе работы над настоящим проектом, открывают возможность создания каталога русских конструкций, соотносимых с определенными значениями целевых слов.

References

1. *Apresjan Ju. D.* (ed.) (2003) Novyj objasnitel'nyj slovar' sinonimov russkogo jazyka [New Explanatory Dictionary of Russian Synonyms]. Moscow.
2. *Apresjan Ju. D.* (ed.) (2010) Prospekt aktivnogo slovarja russkogo jazyka. [The Prospect of Active Dictionary of the Russian Language]. Moscow.
3. *Bolshakova E. I., Baeva N. V., Bordachenkova E. A., Vasil'eva N.E., Morozov S. S.* Leksiko-sintaksicheskie shablony v zadachah avtomaticheskoj obrabotki teksta [Lexico-Syntactic Patterns for Automatic Text Processing]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2007» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2007»]. Moscow, 2007.
4. *Borisova E. G.* (1995) Kollokatsii. Chto eto takoje i kak ih izuchat? [Collocations. What is it and how should we study them?]. Moscow.
5. *Dąbrowska E.* (2004) Language, Mind and Brain: Some Psychological and Neurological Constraints on Theories of Grammar. Edinburgh University Press, Edinburgh and Georgetown University Press, Georgetown.
6. *Fillmore Ch. J.* (1988) The Mechanisms of Construction Grammar. Proceedings of the Berkeley Linguistic Society. Vol. 14.
7. *Gel'buh A. F., Sidorov G. O., èrnandes-Rubio è., Chubukova M. V.* Slovari sochetaemosti slov: kakoj metod sostavlenija luche? [Dictionaries of Word Co-occurrence: Which Way of Building is Better?]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2004» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2004»]. Moscow, 2004.
8. *Goldberg A. E.* (1995) Constructions. A Construction Grammar Approach to Argument Structure. Chicago, IL/London: University of Chicago Press.
9. *Goldberg A. E.* (2006) Constructions at Work: the Nature of Generalization in Language. Oxford: Oxford University Press.
10. *Iordanskaja L. N., Melchuk I. A.* (2007) Smysl i sochetajemost v slovare [Sense and Co-occurrence in a Dictionary]. Moscow.
11. *Jagunova E. V., Pivovarova L. M.* (2011) Ot kollokatsij k konstruktsijam [From Collocations to Constructions], Russkij jazyk: konstruktsionnye i leksiko-semanticheskie podhody [Russian: Constructional and Lexical-Semantic Approaches]. Saint-Petersburg.
12. *Kustova G. I.* Konstruktsii s abstraktnymi sushchestvitel'nymi i ih otrazhenie v èlektronnom slovare [Constructions with Abstract Nouns in an Electronic Database]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog — 2011» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2011»]. Moscow, 2011.
13. *Kuznetsova Ju. L.* (2007) Grammatika konstruktsij. Obzor [Construction Grammar. State of the Art], Nauchno-tehnicheskaja informatsija [Science and Technical Information], Serija 2, № 4.

14. *Lyashevskaya O. N., Kuznetsova Ju. L.* Russkij FrejmNet: k zadache sozdanija korpusnogo slovarja konstruktsij [Russian FrameNet: Towards a Corpus-Based Dictionary of Constructions]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog — 2009» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2009»]. Moscow, 2009.
15. *Lyashevskaya O., Mitrofanova O., Grachkova M., Romanov S., Shimorina A., Shurygina A.* Automatic Word Sense Disambiguation and Construction Identification Based on Corpus Multilevel Annotation. Text, Speech and Dialogue. Proceedings of the 14th International Conference TSD 2011, Pilsen, Czech Republic, September 1–5, 2011. Springer-Verlag, 2011.
16. *Manning C., Schütze H.* (2002) Collocations, in Foundations of Statistical NLP.
17. *Mel'chuk I. A.* (1999/1974) Opyt teorii lingvisticheskih modelej «Smysl <=> Tekst» [On the Theory of Linguistic Models «Sense <=> Text»]. Moscow.
18. *Mel'chuk, Igor A.* (2007) Lexical Functions, H. Burger, D. Dobrovol'skij, P. Kühn & N. Norrick (eds.), Phraseology. An International Handbook of Contemporary Research, 119–213. Berlin/New York: W. de Gruyter.
19. *Mitrofanova O., Lyashevskaya O.* Disambiguation of Taxonomy Markers in Context: Russian Nouns. 17th Nordic Conference of Computational Linguistics NODALIDA — 2009, Odense, Denmark, May 14–16, 2009.
20. *Mitrofanova O. A., Belik V. V., Kadina V. V.* Korpusnoe issledovanie sochetaemostnyh predpochtenij chastotnyh leksem russkogo jazyka [Corpus Analysis of Selectional Preferences of Frequent Words in Russian]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog — 2008» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2008»]. Moscow, 2008.
21. *Mitrofanova O. A., Lyashevskaja O. N., Panicheva P. V.* èksperimenty po statisticheskomu razresheniju leksiko-semanticheskoj neodnoznachnosti russkih imen sushchestvitel'nyh v korpuse [Experiments on Statistical Word Sense Disambiguation of Russian Nouns in the Corpus]. Trudy mezhdunarodnoj konferentsii «Korpusnaja lingvistika–2008» [Proceedings of the International conference «Corpus linguistics — 2008»]. Saint-Petersburg, 2008.
22. *Oskol'skaya S. A, Say S. S.* (2010) Kruglyj stol "Russkij jazyk: konstruktsionnyje i leksiko-semanticheskie podhody" [Workshop on Constructional and Lexico-Grammatical Approaches to Russian, 2009], Voprosy jazykoznanija [Linguistic Inquiries], Vol. 1.
23. *Ovsjannikova M. A., Oskol'skaya S. A, Say S. S.* (2011) Russkij jazyk: konstruktsionnyje i leksiko-semanticheskie podhody [Constructional and Lexico-Grammatical Approaches to Russian], Voprosy jazykoznanija [Linguistic Inquiries], Vol. 5.
24. *Proceedings* of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics. Los Angeles, CA, 2010.
25. *Rakhilina E. V.* (ed.). (2010) Lingvistika konstruktsij [Construction Linguistics]. Moscow.
26. *Rakhilina E. V.,Kobritsov B. P.,Kustova G. I.,Lyashevskaja O. N.,Shemanaeva O. Ju.* Mnogoznachnost' kak prikladnaja problema: leksiko-semanticheskaja razmetka

v Natsional'nom korpuse russkogo jazyka [Semantic Ambiguity as an Application-Oriented Problem: Word Class Tagging in the RNC]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog — 2006» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2006»]. Moscow, 2006.

27. *Sahlgren M., Knutsson O.* (2009) Workshop on Extracting and Using Constructions in NLP. NODALIDA'09. SICS Technical Report T2009:10.
28. *Shimorina A., Grachkova M.* Identification of Context Markers for Russian Nouns. 18th Nordic Conference of Computational Linguistics NODALIDA — 2011, Riga, Latvia, May 11–13, 2011.
29. *TKS* — Mel'chuk I. A., Zholkovskij A. K. i dr. (1984) Tolkovo-kombinatornyj slovar' sovremennogo russkogo jazyka. Opyty semantiko-sintaksicheskogo opisanija russkoj leksiki [Explanatory Combinatorial Dictionary og Modern Russian. Semantic-Syntactic Studies of Russian Vocabulary]. Vienna: Wiener Slavistischer Almanach.
30. *Tomasello M.* (2003) Constructing a Language: A Usage-Based Approach to Child Language Acquisition. Cambridge, MA: Harvard University Press.
31. *Wible D., Tsao N.-L.* StringNet as a Computational Resource for Discovering and Investigating Linguistic Constructions. Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics, Los Angeles, CA, 2010.
32. *Zakharov V. P., Khokhlova M. V.* Analiz effektivnosti statisticheskih metodov vyjavlenija kollokatsij v tekstah na russkom jazyke [The Study of Effectiveness of Statistical Measures for Collocation Extraction on Russian Texts]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2010» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2010»]. Moscow, 2010.