Опыт автоматической каталогизации русских именных конструкций (на материале НКРЯ)♣

О.А. Митрофанова[1], О.Н. Ляшевская[2], М.А. Грачкова[1],
С.В. Романов[3], А.С. Шиморина[4], А.С. Шурыгина[5]

[1]Санкт-Петербургский государственный университет, [2]Высшая школа экономики,
[3]ЗАО «Интернет-Проекты», [4]Институт лингвистических исследований РАН,
[5]Российский государственный гуманитарный университет
alkonost-om@yandex.ru, olesar@gmail.com, maaag86@mail.ru,
complefor@rambler.ru, shinas@yandex.ru, sanyana@gmail.com

Аннотация

Цель проекта – автоматическое построение каталога русских конструкций на базе Национального корпуса русского языка (НКРЯ, http://www.ruscorpora.ru/). В докладе анализируется один из важнейших аспектов проекта, а именно, построение шаблонов именных конструкций на основе контекстных маркеров значений целевых слов. Обсуждается компьютерный инструмент автоматического выделения конструкций, результаты экспериментов и их верификация.

1. Введение

Цель обсуждаемого в докладе проекта – автоматическое построение каталога русских конструкций на базе Национального корпуса русского языка (НКРЯ, http://www.ruscorpora.ru/). Исследование проводится коллективами НКРЯ и кафедры математической лингвистики СПбГУ [1, 2, 3].

Проблема автоматического выделения конструкций сегодня привлекает внимание многих ученых [4, 5]. Успешно решается задача извлечения *n*-грамм (коллокаций, неоднословных целостностей), вместе с тем, конструкции, имеющие нестандартную семантико-синтаксическую организацию, представляют сложность в автоматической обработке текста. Как в отечественной, так и в зарубежной практике уже реализованы проекты, направленные на формализацию лексико-синтаксических связей единиц текста (извлечение лексико-синтаксических шаблонов, автоматическое построение словарей сочетаемости, Word Sketches для русского языка [6] и т.д.). Однако, русскоязычные ресурсы, отражающие сочетаемость и рамки валентностей, недостаточно разработаны. Создание систем типа FrameNet или PropBank для русского языка способствовало бы выполнению задачи каталогизации конструкций [7].

2. Объект исследования

Конструкция в нашем проекте трактуется как сочетание целевого слова и контекстных маркеров его значения, характеризующееся частотностью и устойчивостью. Данное определение конструкций согласуется с основными идеями Грамматики конструкций [8, 9, 10]. 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, подробнее см.: http://www.ruscorpora.ru/corpora-sem.html).

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

3. Лингвистические источники

Источником данных об именных конструкциях служат выборки контекстов из НКРЯ, крупнейшего русскоязычного корпуса с многоярусной разметкой. В центре внимания находятся

(1) отдельные существительные из различных лексико-семантических групп: *дом, вид, орган, лук, глава*…; (2) целостные лексико-семантические группы существительных – названий инструментов (*бритва, веник, весло, карандаш, коса*…), обозначений речевых действий (*дискуссия, комплимент, обращение, обсуждение, ответ*…) и т.д. Разметка значений слов в контекстах НКРЯ проводилась на основе Семантического словаря НКРЯ.

Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2012), Санкт-Петербург, Россия, 2012.

4. Программное обеспечение экспериментов

Компьютерный инструмент выделения конструкций, разработанный на языке Python, 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 (см., например, табл. 2). Эти списки проходят постредактирование, в ходе которого отсеиваются сочетания, малоинформативные для определения значения целевого слова в контексте.

5. Результаты экспериментов

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. Ниже приведены примеры двучленных конструкций целевых слов, выражающих инструментальное значение **r:concr t:tool:instr:**

(1) A+S / S+A:

– тип инструмента: **r:rel der:s:** *механический*, электрический + БРИТВА и т.д.;

– сфера применения инструмента: **r:rel der:s dt:space:** *садовый* + НОЖНИЦЫ и т.д.;

(2) V+S / S+V:

– действия, производимые при помощи инструмента: ВЕНИК + **d:pref der:v | t:move ca:caus d:pref der:v:** *подметать* и т.д.;

– действия, производимые применительно к инструменту: весло + **d:root:** *бросать* и т.д.;

(3) S + S:

– совместно используемые инструменты: МЕТЛА + **r:concr t:tool:instr:** *лопата, швабра, щетка, веник* и т.д.

– части инструментов: **r:concr t:thing pt:residpart pc:X der:v:** *огрызок* + КАРАНДАШ и т.д.

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 (см. пример в табл. 2).

Таблица 1. Пример обработки данных о конструкциях со словом *ножницы*

Левосторонние соседи (позиция -1)			**Параметры обработки данных**	**Правосторонние соседи (позиция +1)**		
gr	sem	lex		gr	sem	lex
V	**t:impact ca:caus d:root:**	резать(10) кромсать(1) стричь(4)	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	**V**	**t:impact ca:caus d:pref:**	нарезать(2) срезать(1) перерезать(3) искромсать(1)
	t:poss ca:caus:	взять(17)			**t:impact ca:caus d:pref der:v:**	вырезать(4) надрезать(3) обрезать(1)
	t:impact ca:caus d:pref der:v:	вырезать(9) обрезать(4) надрезать(1) обстригать(1)				
Adj	**r:rel der:s dt:space:**	садовый(19)		**Adj**	**r:rel:**	что(3) известный(1) дикий(1) лишний(1) обыкновенный(1)
Pron	**r:dem:**	этот(11) такой(1) тот(1)		**Pron**	**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)
					r:pers:	она(2) я(3) он(1) они(2) мы(1) вы(1)
...	...	...		...	…	…

Таблица 2. Устойчивые сочетания в составе конструкций со словом *ножницы*

НОЖНИЦЫ: режущий инструмент из двух раздвигающихся лезвий с кольцеобразными ручками (**r:concr t:tool:instr der:s**)		
Конструкции	Сервис поиска биграмм С. А. Шарова	Словари
r:rel der:s dt:space: *садовый* + НОЖНИЦЫ	**садовый** LL = 40.50	**Садовые** ножницы.
r:rel der:s: *маникюрный*, *хирургический* + НОЖНИЦЫ	**маникюрный** LL = 49.47	**Маникюрные** ножницы.

6. Заключение

Исследование показывает, что проблема автоматического выделения конструкций имеет практическое решение:

(1) разработан инструмент автоматического выделения конструкций;

(2) разработана и апробирована методика анализа данных о конструкциях, предложены принципы их систематизации;

(3) проведены пилотные эксперименты по выделению конструкций для русских существительных;

(4) проведена верификация экспериментальных данных.

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

Литература

[1] 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.

[2] 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.

[3] 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.

[4] Sahlgren M., Knutsson O. Workshop on Extracting and Using Constructions in NLP. NODALIDA’09. SICS Technical Report T2009:10. 2009.

[5] Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics. Los Angeles, CA, 2010.

[6] Захаров В.П., Хохлова М.В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке// Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог – 2010». М., 2010.

[7] Ляшевская О.Н., Кузнецова Ю.Л. Русский Фреймнет: к задаче создания корпусного словаря конструкций // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог – 2009». М., 2009.

[8] Fillmore Ch.J. The Mechanisms of Construction Grammar. Proceedings of the Berkeley Linguistic Society. Vol. 14. 1988.

[9] Кузнецова Ю.Л. Грамматика конструкций. Обзор // Научно-техническая информация. Серия 2, № 4, 2007.

[10] Лингвистика конструкций / Под ред. Е.В. Рахилиной. М., 2010.

On Automatic Catalogization of Russian Nominal Constructions (Basen on RNC)

O.A. Mitrofanova, O.N. Lashevskaya, M.A. Grachkova, S.V. Romanov, A.S. Shimorina, A.S. Shurygina

The purpose of the project is automatic construction identification performed on the Russian National Corpus (RNC, http://www.ruscorpora.ru/). The talk is focused on the extraction of patterns for nominal constructions taking into account context markers of target words. We also discuss the toolbox developed for automatic construction identification, experimental results and their verification.

♣ 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».