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

Ляшевская О. Н. (olesar@gmail.com),
НИУ ВШЭ, ИРЯ РАН, Москва, Россия

Паничева П. В. (ppolin86@gmail.com),
EPAM Systems, Россия

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

Инструмент для лексикографа: заготовка для последующего ручного отбора и редактирования

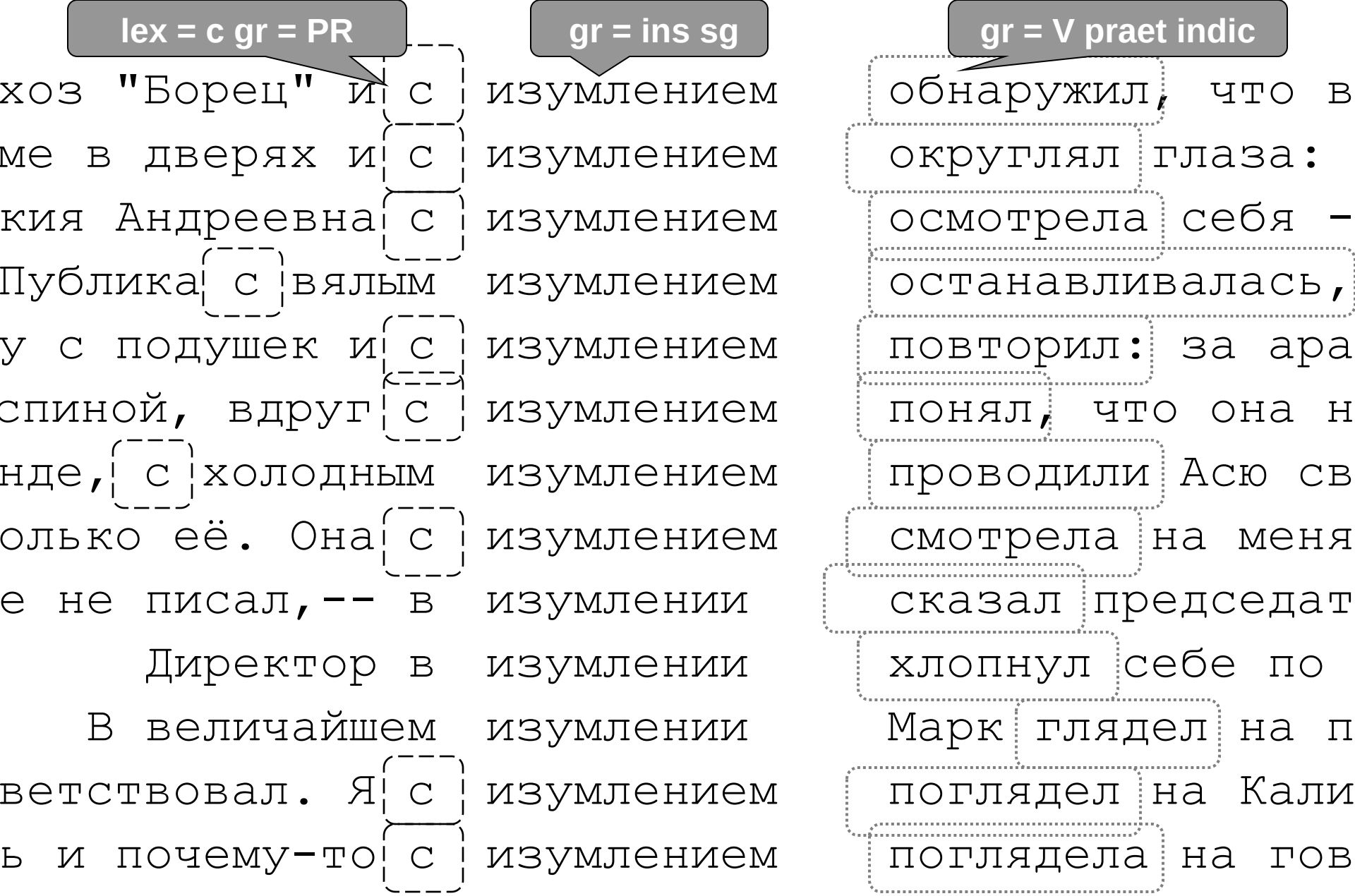
- **Каковы типичные контексты, в которых употребляется слово?**
- **Кластеризация корпусных данных, более "умная", чем списки коллокаций**
- **Задача 1: увидеть за паттернами "лингвистическую" структуру, определить ее в терминах грамматических признаков и лексических единиц и классов**
- **Имитируем работу лингвиста?**
- **Задача 2: соотнести получившиеся "конструкции" с разными значениями слова**
- **Задача 3: представить конструкции визуально**

История проекта

Сотрудничество Национального корпуса русского языка (НКРЯ, http://ruscorpora.ru) и кафедры математической лингвистики СПбГУ

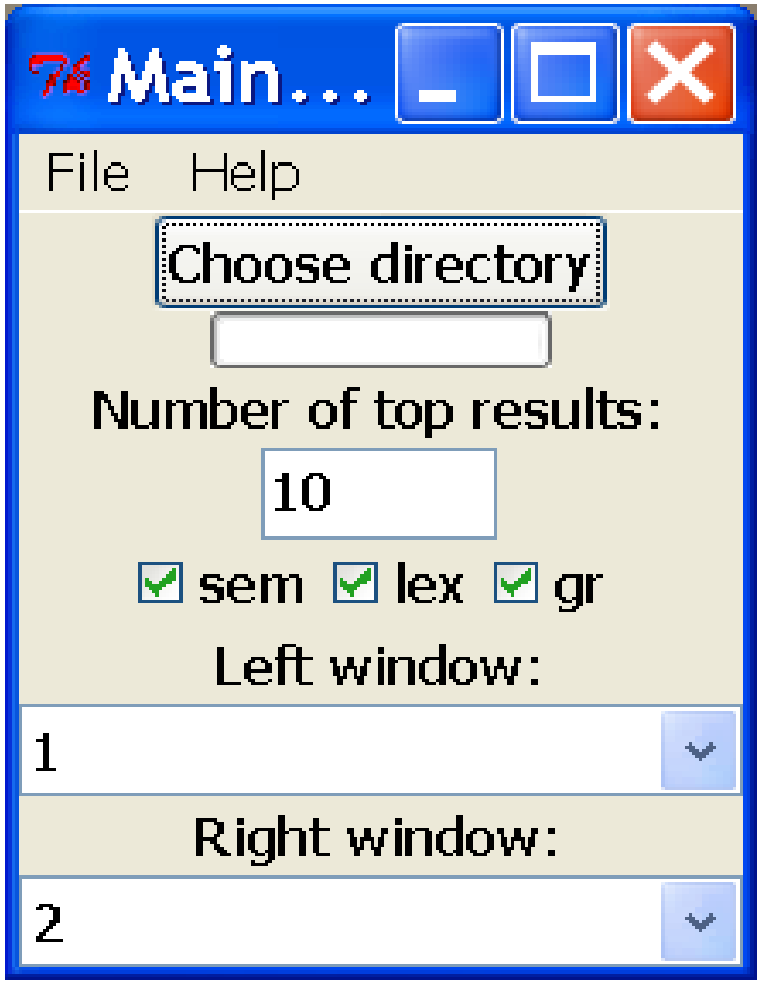
Данные

- имена **речевых действий** (*дискуссия, комплимент, обращение, обсуждение, ответ* и т.д.),
- названия **эмоций** (*апатия, благодарность, грусть, гнев, любовь* и т.д.),
- названия **инструментов** (*бритва, веник, весло, карандаш, коса* и т.д.)



Кластеризация конструкций по корпусным данным

Программы выделения конструкций: пользовательский интерфейс



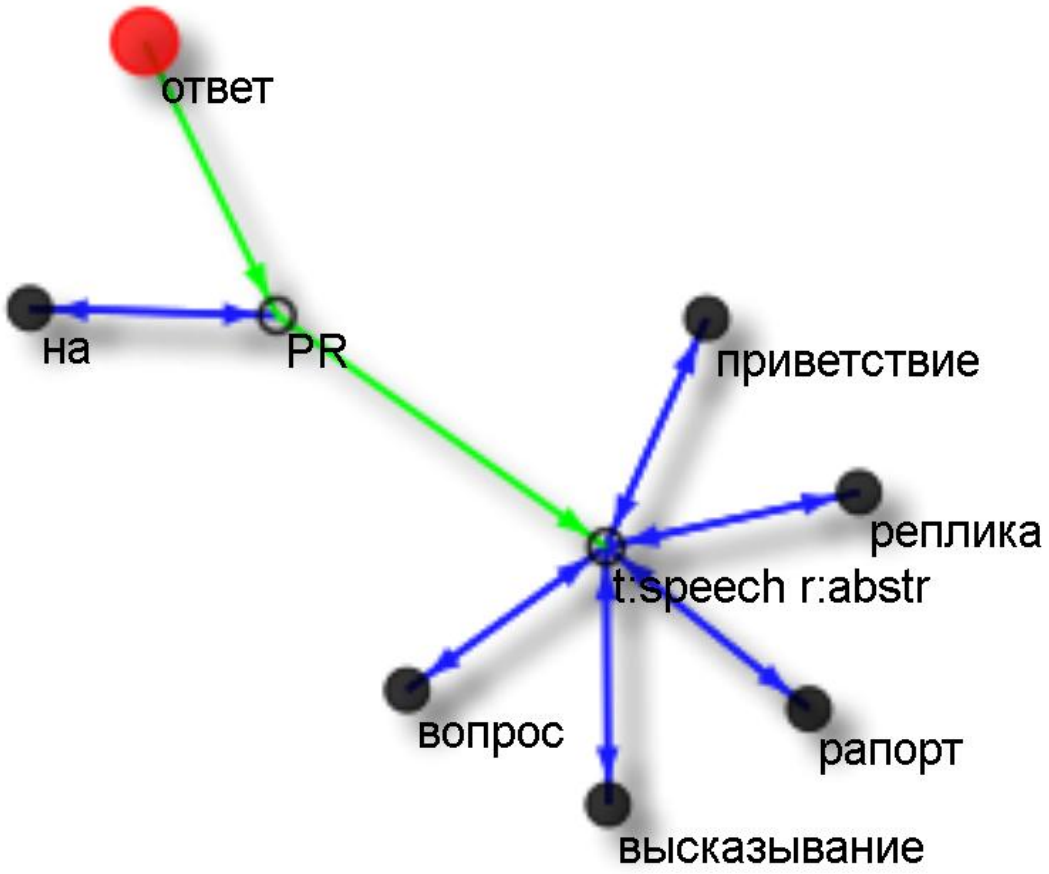
Конструкция – это комбинация целевого слова и слотов, заполняемых регулярными контекстными соседями, среди которых могут быть ***lex*** – леммы, ***gr*** – грамматические, ***sem*** – лексико-семантические и т. п. признаки.

Конструкция – это абстрактный шаблон, предполагающий **лексикализацию**, т.е. различные реализации в виде комбинаций лемм/словоформ

V|*дать, найти, предложить…* ОТВЕТ + PR|*на* + speech r:abstr|*вопрос*,

r:qual|*простой, неоднозначный…* + ОТВЕТ, ОТВЕТ + t:hum r:concr|*академикам, мудрецам, отцу...*

ОТВЕТ* + PR|*на* + t:speech r:abstr| *приветствие, опрос, высказывание, рапорт, реплика



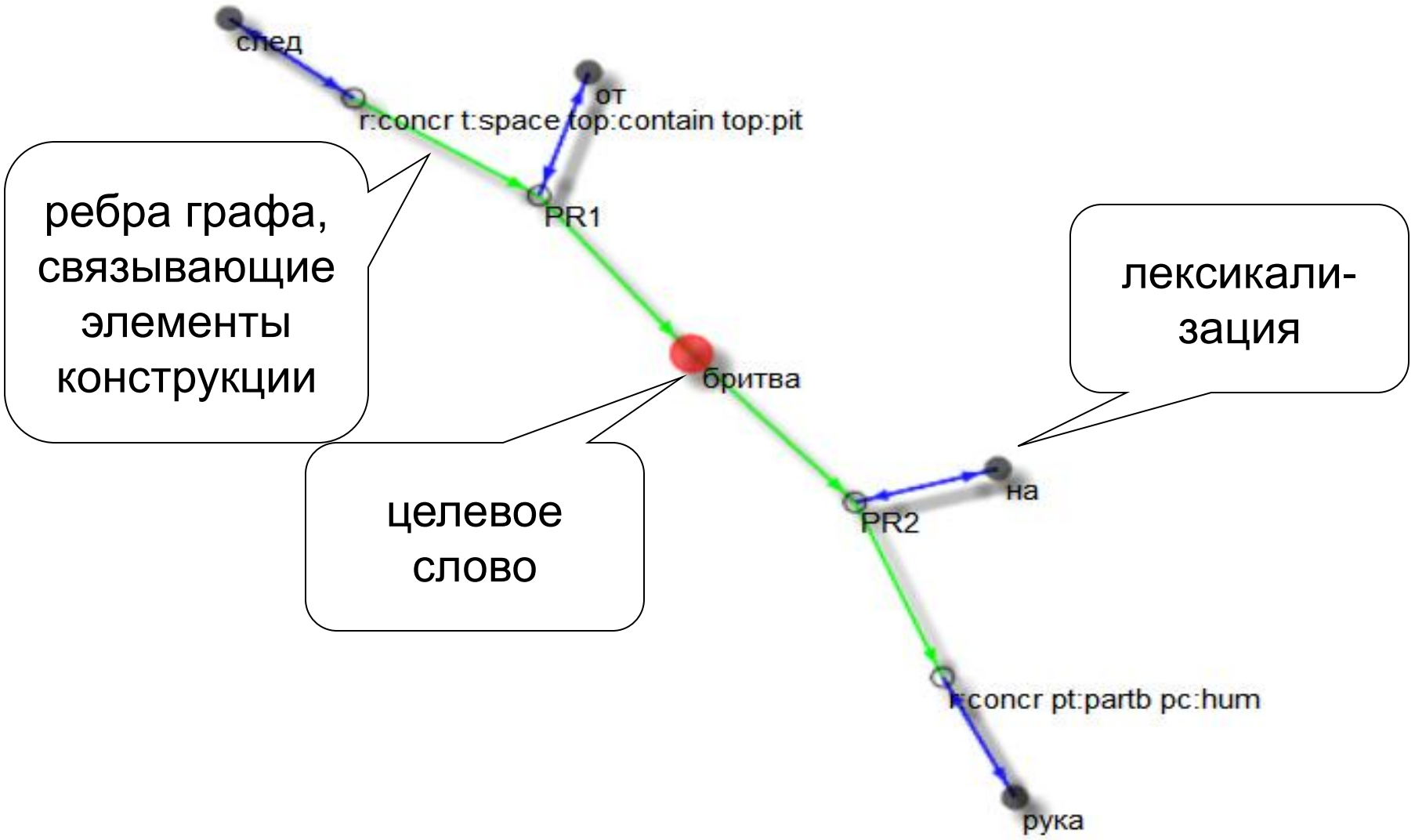
Визуализация данных о конструкциях

Python-библиотека pattern.graph
(http://www.clips.ua.ac.be/pages/pattern-graph)

2 этапа:

1) парсинг строки конструкции и выявление ее главных и второстепенных элементов с сохранением порядка
2) рисование графа

r:concr t:space top:contain top:pit| *след* + PR| *от* + *БРИТВА* + PR| *на* + r:concr pt:partb pc:hum| *рука*



Перспективы

- визуализация трех слоев разметки (леммы, грамматические теги, лексико-семантические теги)
- отражение факультативных элементов в конструкции
- динамическая визуализация – когда конструкции "вкладываются" друг в друга и имеют много лексических вариантов реализации
- представление "наслаивания" нескольких конструкций друг на друга в контексте
- оценка: сопоставление выделенных наборов лексических конструкций с наборами, который мог бы выделить лексикограф на тех же данных.