Гатауллин Р.Р.* Гильмуллин Р.А.*

ПРОГРАММНЫЙ ИНСТРУМЕНТАРИЙ ДЛЯ РАЗРЕШЕНИЯ МОРФОЛОГИЧЕСКОЙ МНОГОЗНАЧНОСТИ В ТАТАРСКОМ ЯЗЫКЕ

** Казанский федеральный университет,*
Научно-исследовательский институт «Прикладная семиотика» АН РТ,
г. Казань, Республика Татарстан

Морфологическая многозначность

- Пример функциональной омонимии из русского языка

  “Три яблока”

  “Три сильнее!”

- Пример функциональной омонимии из татарского языка:

  “Тезен тезгә куйды.”
  (Поставил колено на колено)

  “Сафка тезен!” (Встать в строй!)

Статистика по корпусу

- По подкорпусу в 21млн словоупотреблений:
- ~ 31% словоформ имеют более чем один морфологический разбор
- ~25% словоформ имеют два разбора
- Более 10.000 типов морфологической многозначности

  “алма” → “яблоко | не бери” = N+Sg | V+Neg+Imp_Sg

  “ат” → “конь | стреляй” = N+Sg | V+Imp_Sg
- (теоретически) можно свести к 420 классам

Методы

- Контекстные методы
- Статистико-вероятностные методы
- Гибридные методы

Контекстный метод

Преимущества:

- точность выше, чем у других методов
- не требует размеченного корпуса

Трудность:

- требует тщательного лингвистического анализа каждого типа

Статистико-вероятностные методы

Преимущества:

не требуют такого тщательного анализа и разработки самих правил

Трудность:

требуется размеченный корпус текстов

Гибридный метод

Контекстный метод + статистико-вероятностный метод

Преимущества:

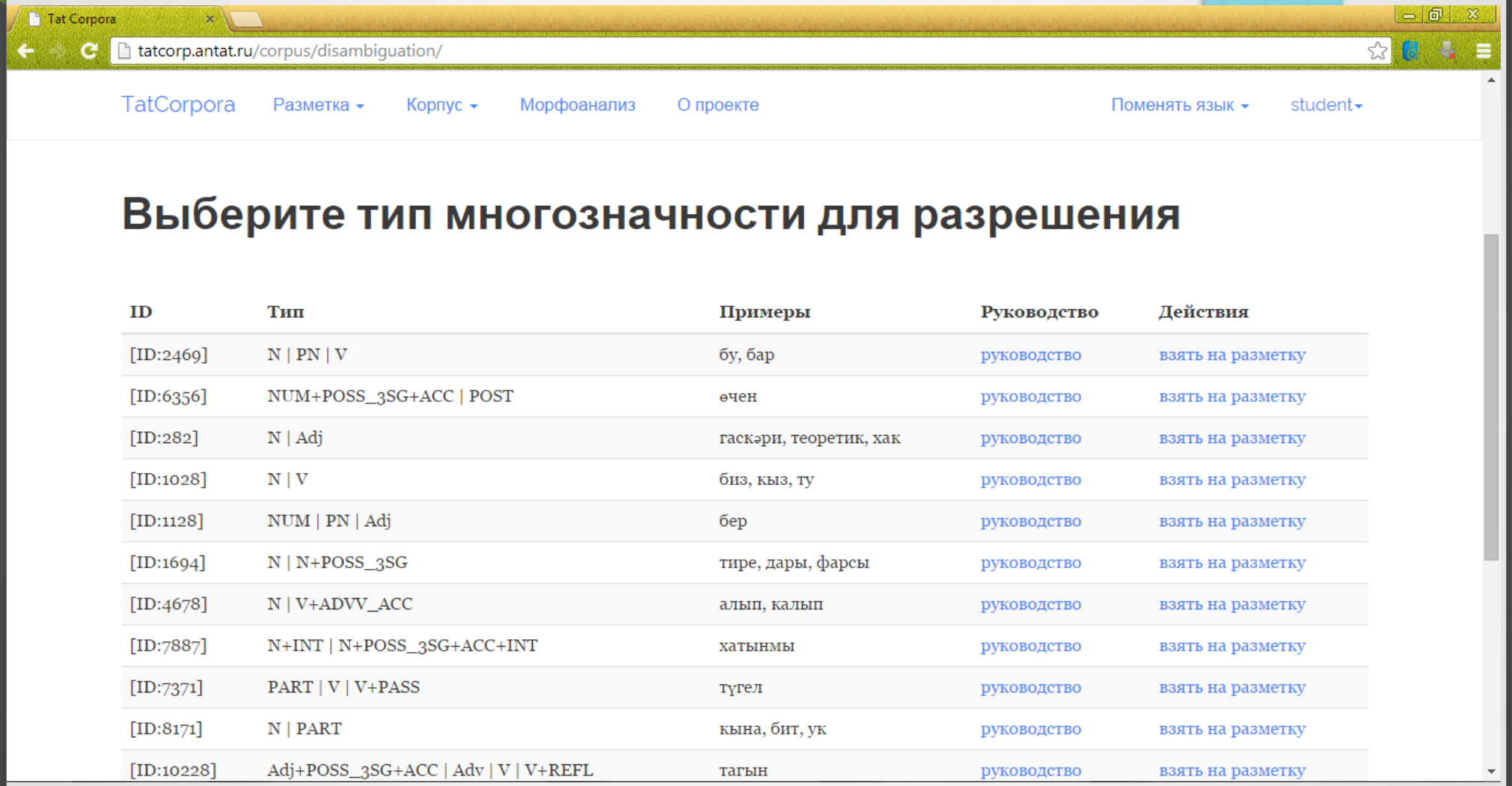
- преимущества одного метода перекрывают недостатки другого

tatcorp.antat.ru

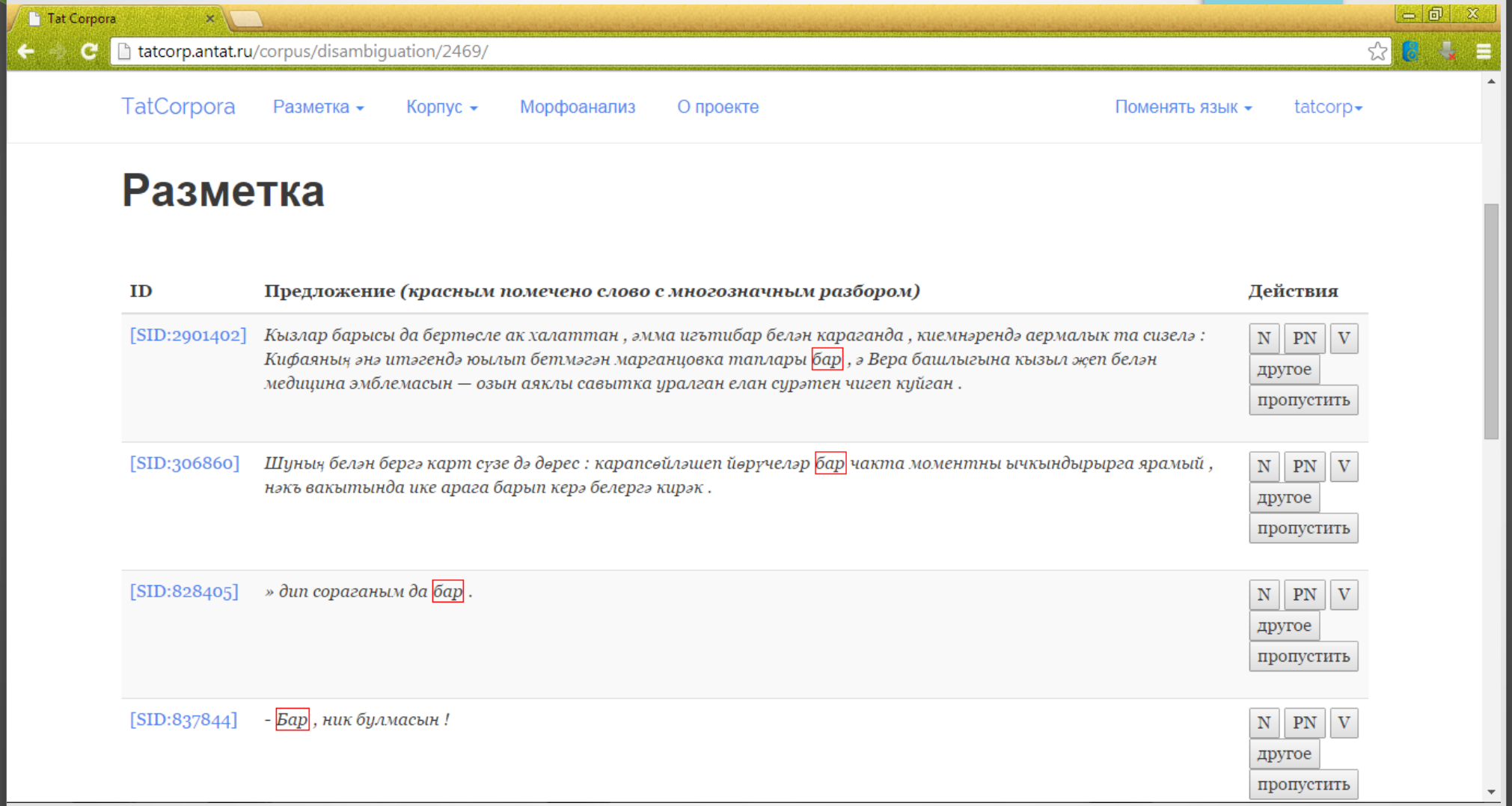
Решаемые задачи:

- Ручное снятие морфологической многозначности
- Разработка и тестирование контекстных правил разрешения морфологической многозначности
- Исправление ошибок разметки и морфоанализатора

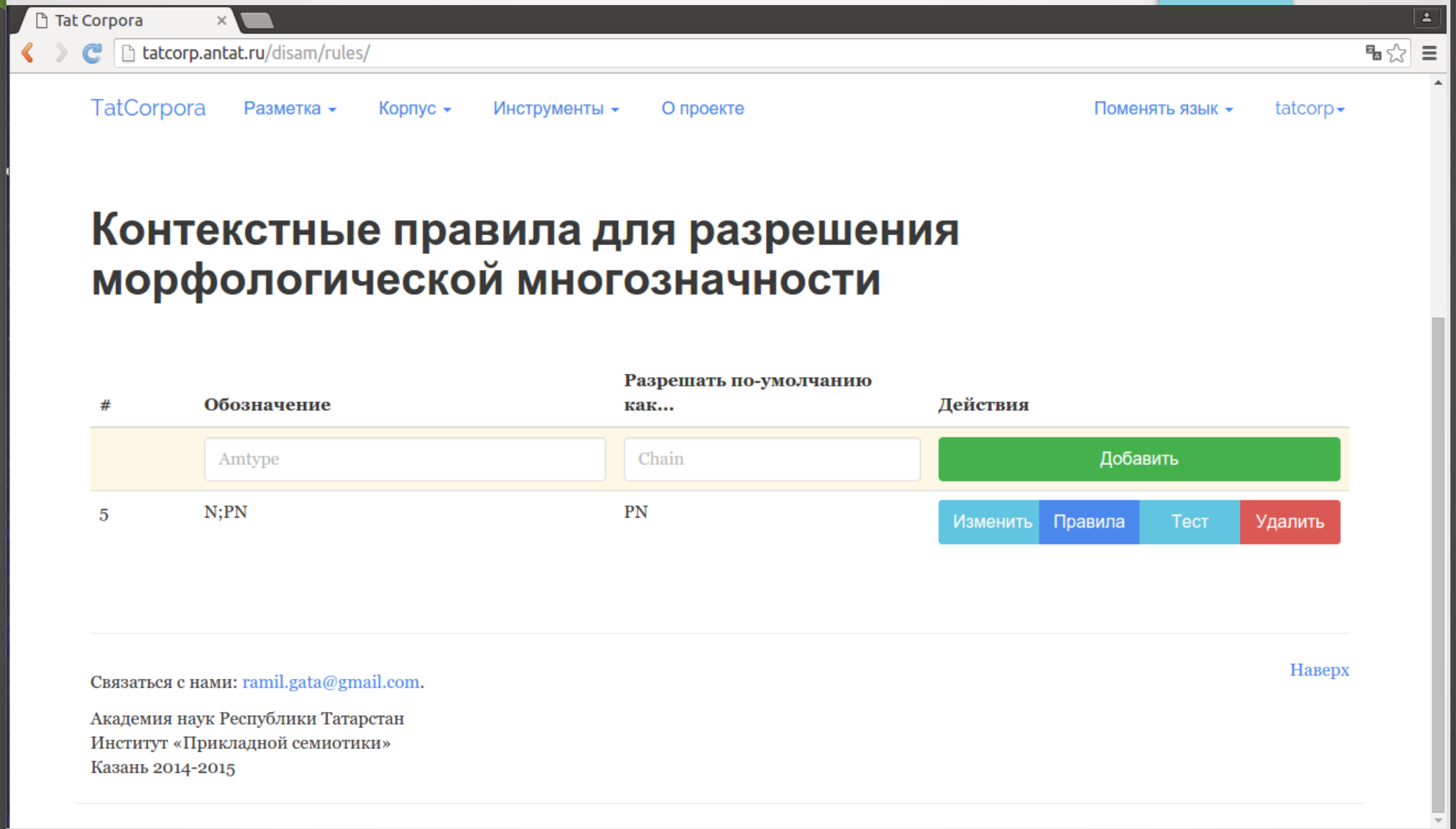
tatcorp.antat.ru/corpus/disambiguation



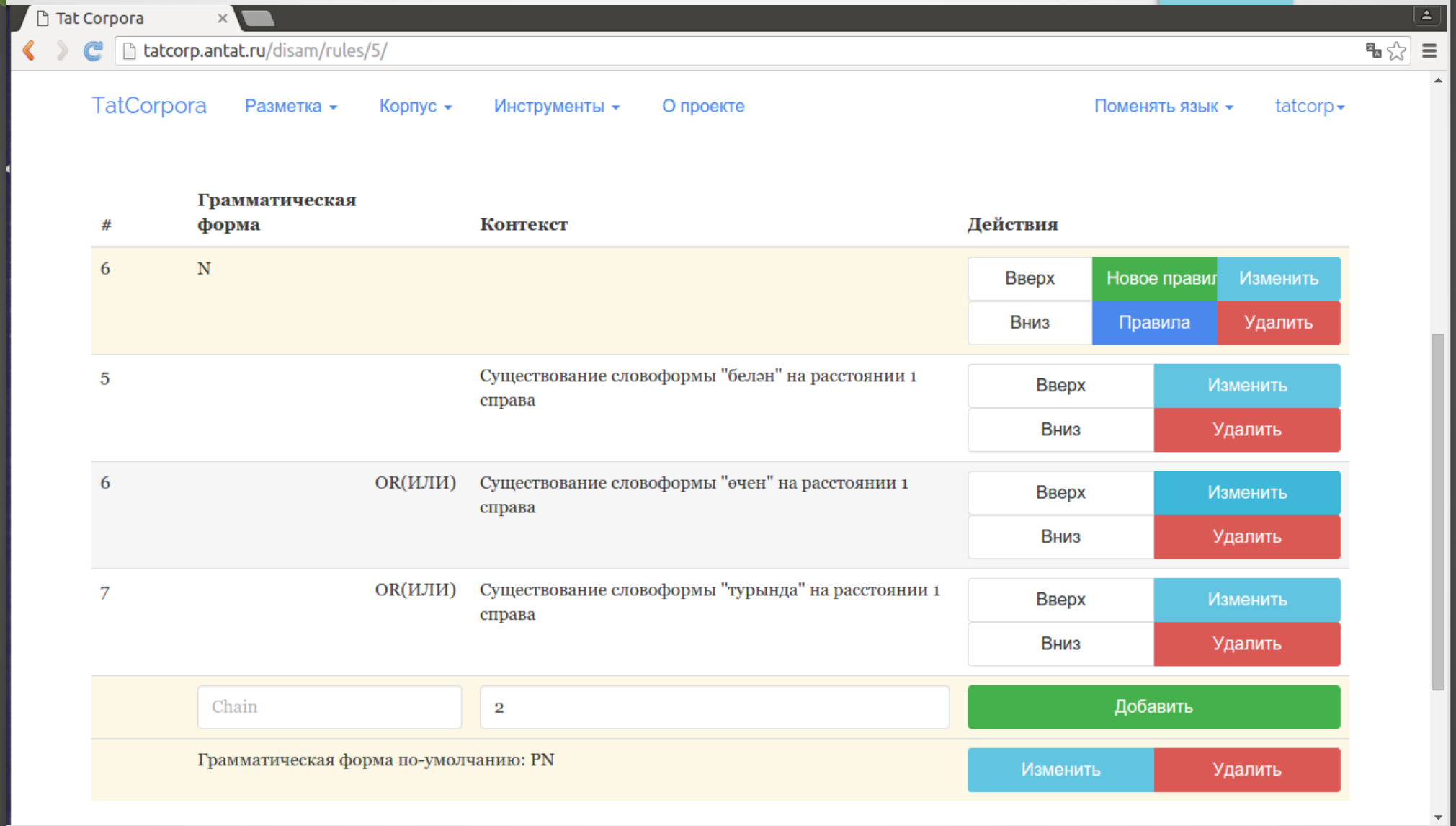
tatcorp.antat.ru/corpus/disambiguation/2469



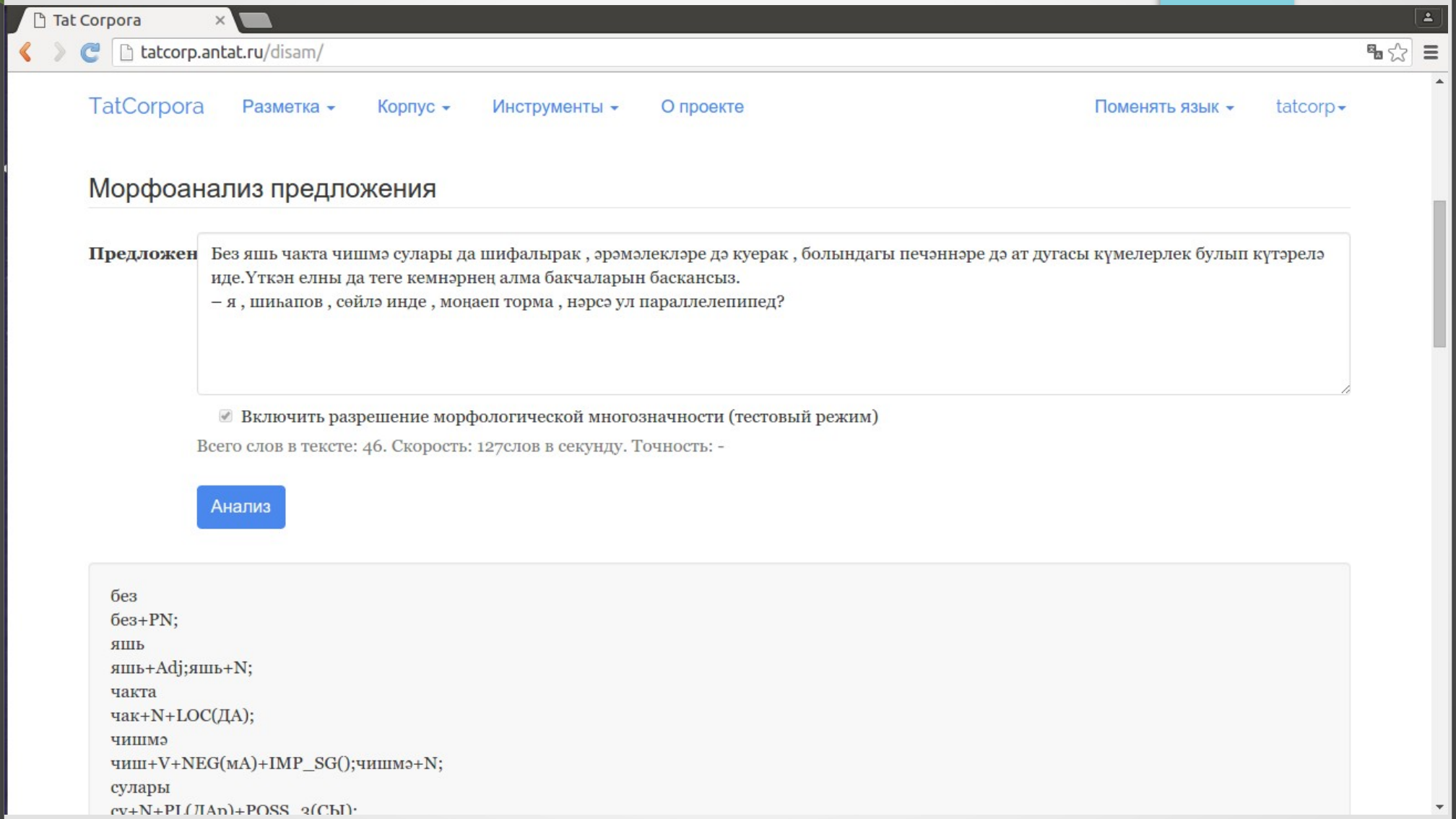
tatcorp.antat.ru/disam/rules



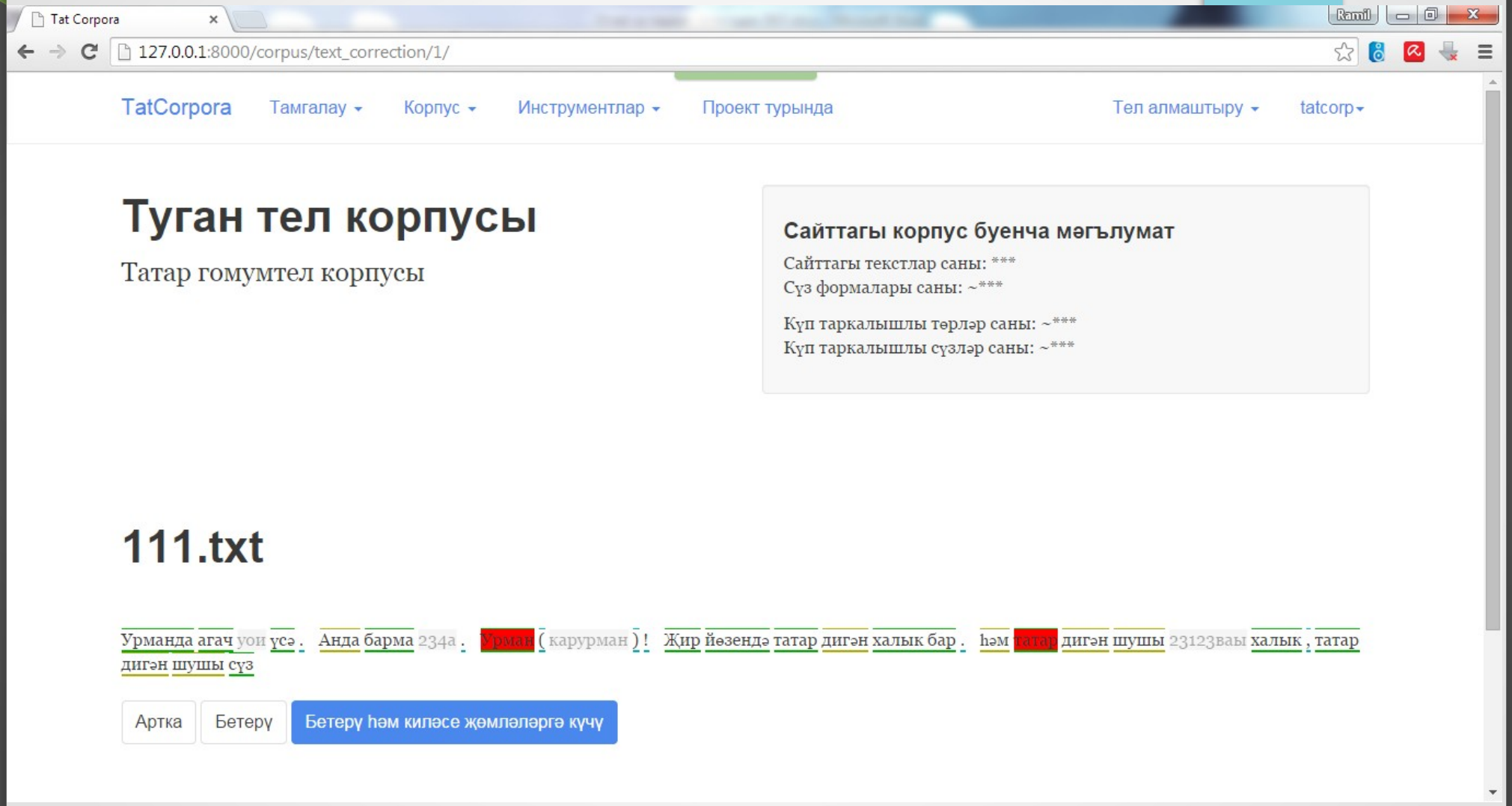
tatcorp.antat.ru/disam/rules/5



tatcorp.antat.ru/disam



tatcorp.antat.ru/text_correction/1



Технологии

- Python 3.3
- Фреймворк Django 1.7
- СУБД PostgreSQL

Спасибо за внимание!

Гатауллин Р.Р.* Гильмуллин Р.А.*

ПРОГРАММНЫЙ ИНСТРУМЕНТАРИЙ ДЛЯ РАЗРЕШЕНИЯ МОРФОЛОГИЧЕСКОЙ МНОГОЗНАЧНОСТИ В ТАТАРСКОМ ЯЗЫКЕ

** Казанский федеральный университет,*
Научно-исследовательский институт «Прикладная семиотика» АН РТ,
г. Казань, Республика Татарстан