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:

```
<u>"Тезен</u> тезгә куйды."(Поставил колено на колено)"Сафка <u>тезен</u>!" (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:

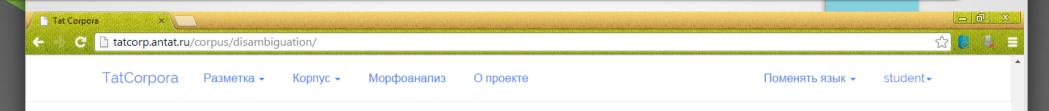
преимущества одного метода перекрывают недостатки другого

tatcorp.antat.ru

Решаемые задачи:

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

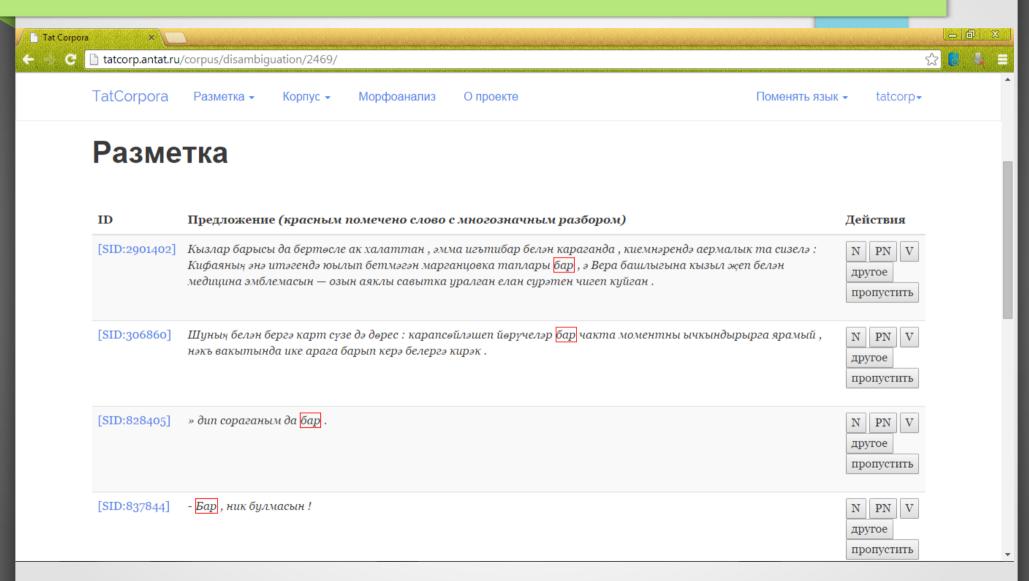
tatcorp.antat.ru/corpus/disambiguation



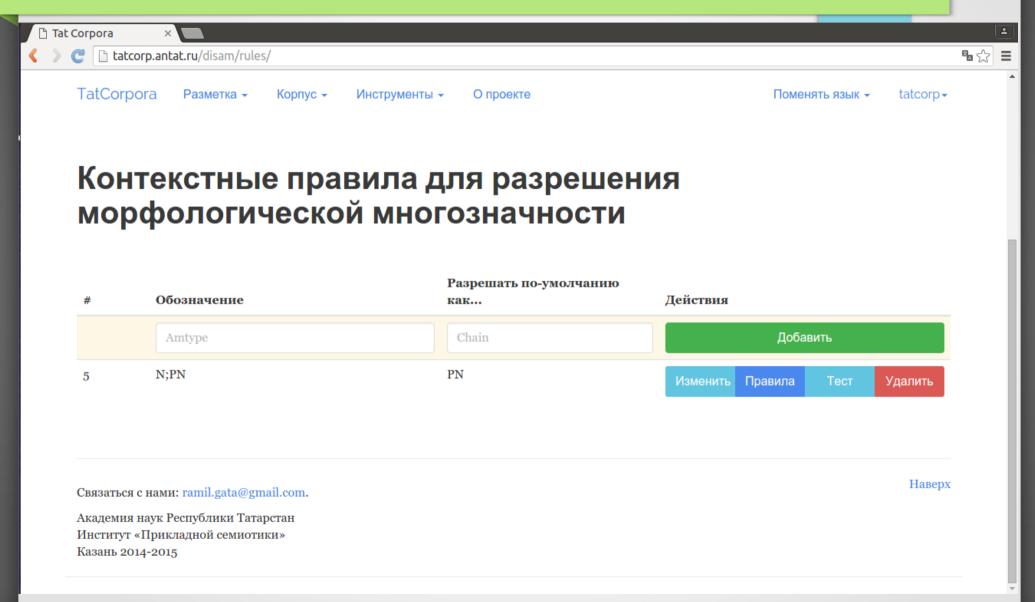
Выберите тип многозначности для разрешения

ID	Тип	Примеры	Руководство	Действия
[ID:2469]	N PN V	бу, бар	руководство	взять на разметку
[ID:6356]	NUM+POSS_3SG+ACC POST	өчен	руководство	взять на разметку
[ID:282]	N Adj	гаскәри, теоретик, хак	руководство	взять на разметку
[ID:1028]	N V	биз, кыз, ту	руководство	взять на разметку
[ID:1128]	NUM PN Adj	бер	руководство	взять на разметку
[ID:1694]	N N+POSS_3SG	тире, дары, фарсы	руководство	взять на разметку
[ID:4678]	N V+ADVV_ACC	алып, калып	руководство	взять на разметку
[ID:7887]	N+INT N+POSS_3SG+ACC+INT	хатынмы	руководство	взять на разметку
[ID:7371]	PART V V+PASS	түгел	руководство	взять на разметку
[ID:8171]	N PART	кына, бит, ук	руководство	взять на разметку
[ID:10228]	Adj+POSS_3SG+ACC Adv V V+REFL	тагын	руководство	взять на разметку

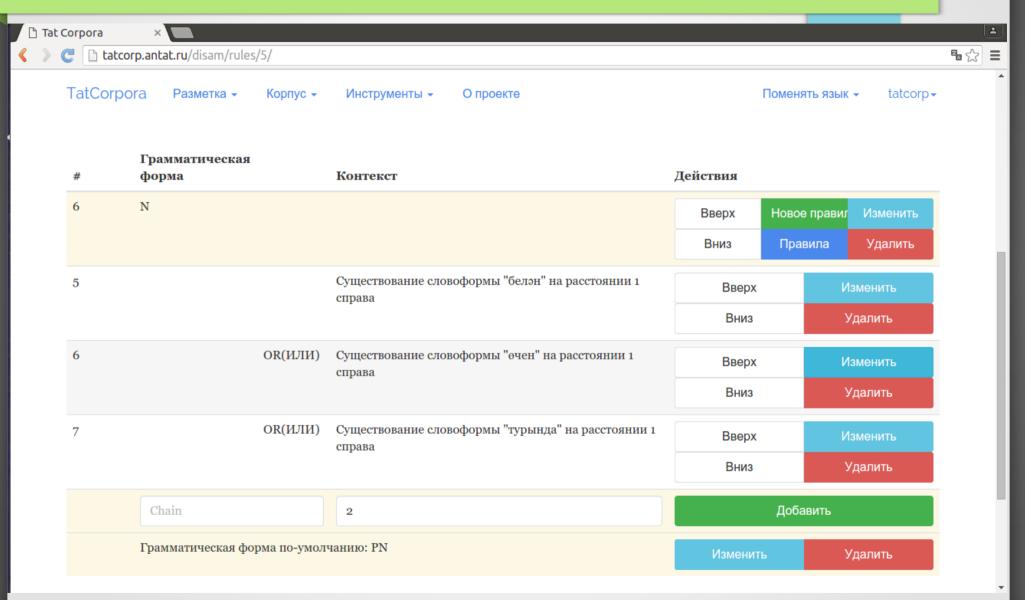
tatcorp.antat.ru/corpus/disambiguation/2469



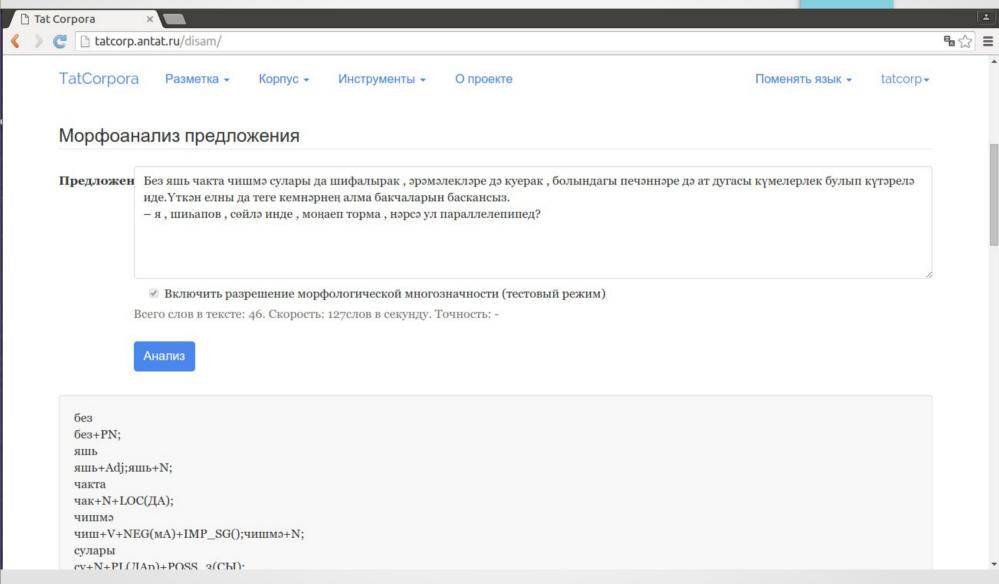
tatcorp.antat.ru/disam/rules



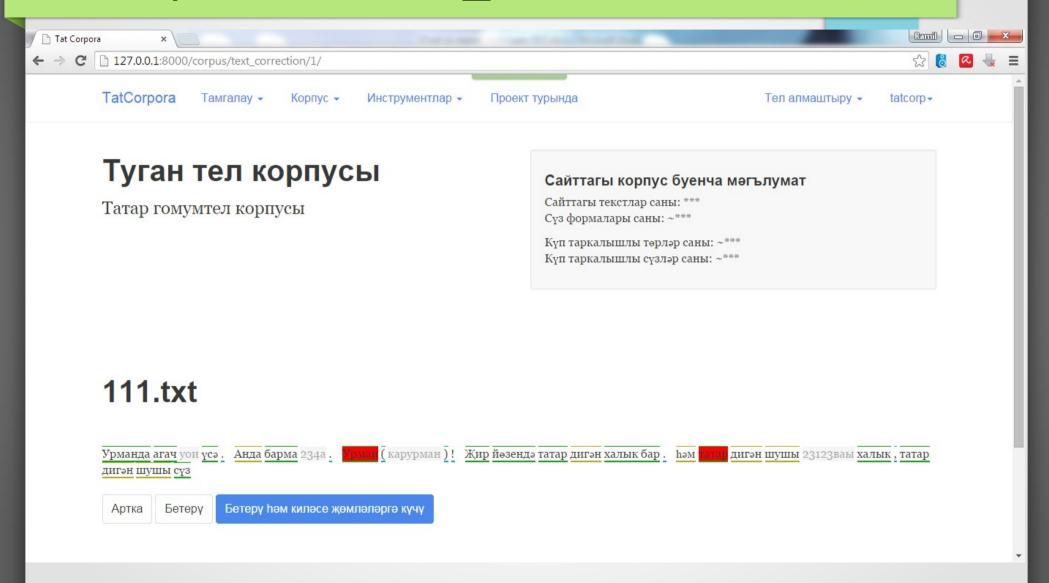
tatcorp.antat.ru/disam/rules/5



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