

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

## Решаемые задачи:

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



# tatcorp.antat.ru/corpus/disambiguation

Tat Corpora

tatcorp.antat.ru/corpus/disambiguation/

Поменять язык ▾ student ▾

Разметка ▾

Корпус ▾

Морфоанализ

О проекте

Выберите тип многозначности для разрешения

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

# tatcorp.antat.ru/corpus/disambiguation/2469

Tat Corpora   Разметка ▾   Корпус ▾   Морфоанализ   О проекте   Поменять язык ▾   tatcorp ▾

## Разметка

ID	Предложение (красным помечено слово с многозначным разбором)	Действия
[SID:2901402]	Кызлар барысы да бертөсле ак халаттан , эмма игътибар белэн караганда , киемнэрэндэ аермалык та сизелэ : Кифаяның энэ итэгэндэ юылып бетмэгэн марганцовка таплары <b>бар</b> , э Вера башлыгына кызыл жеп белэн медицина эмблемасын — озын аяклы савытка уралган елан сурэтен чигеп куйган .	<div>N   PN   V</div> <div>другое</div> <div>пропустить</div>
[SID:306860]	Шуның белэн бергэ карт сүзе дэ дөрес : карасөйлэшеп йөрүчелэр <b>бар</b> чакта моментны ычкындырырга ярамый , нэкъ вакытында ике арага барып керэ белергэ кирэк .	<div>N   PN   V</div> <div>другое</div> <div>пропустить</div>
[SID:828405]	» дип сораганым да <b>бар</b> .	<div>N   PN   V</div> <div>другое</div> <div>пропустить</div>
[SID:837844]	- <b>Бар</b> , ник булмасын !	<div>N   PN   V</div> <div>другое</div> <div>пропустить</div>

# tatcorp.antat.ru/disam/rules

Tat Corpora

tatcorp.antat.ru/disam/rules/

TatCorpora

Разметка

Корпус

Инструменты

О проекте

Поменять язык

tatcorp

## Контекстные правила для разрешения морфологической многозначности

#	Обозначение	Разрешать по-умолчанию как...	Действия
	<input type="text" value="Amtype"/>	<input type="text" value="Chain"/>	<div>Добавить</div>
5	N;PN	PN	<div>Изменить</div> <div>Правила</div> <div>Тест</div> <div>Удалить</div>

Связаться с нами: [ramil.gata@gmail.com](mailto:ramil.gata@gmail.com).

Наверх

Академия наук Республики Татарстан  
Институт «Прикладной семиотики»  
Казань 2014-2015

# tatcorp.antat.ru/disam/rules/5

TatCorpora Разметка Корпус Инструменты О проекте Поменять язык tatcorp

#	Грамматическая форма	Контекст	Действия
6	N		<div>Вверх Новое правило Изменить</div> <div>Вниз Правила Удалить</div>
5		Существование словоформы "белэн" на расстоянии 1 справа	<div>Вверх Изменить</div> <div>Вниз Удалить</div>
6	OR(ИЛИ)	Существование словоформы "өчен" на расстоянии 1 справа	<div>Вверх Изменить</div> <div>Вниз Удалить</div>
7	OR(ИЛИ)	Существование словоформы "турында" на расстоянии 1 справа	<div>Вверх Изменить</div> <div>Вниз Удалить</div>
	<input type="text" value="Chain"/>	<input type="text" value="2"/>	<div>Добавить</div>
	Грамматическая форма по-умолчанию: PN		<div>Изменить Удалить</div>

# tatcorp.antat.ru/disam

Tat Corpora x

tatcorp.antat.ru/disam/

TatCorpora Разметка Корпус Инструменты О проекте Поменять язык tatcorp

## Морфоанализ предложения

**Предложен** Без яшь чакта чишмэ сулары да шифалырак , эрэмэлектлэре дэ куерак , болындагы печәннэре дэ ат дугасы күмелерлек булып күтэрелә иде. Үткән елны да теге кемнэрнең алма бакчаларын баскансыз.  
– я , шиһапов , сөйлә инде , моңаеп торма , нәрсә ул параллелепипед?

☒ Включить разрешение морфологической многозначности (тестовый режим)

Всего слов в тексте: 46. Скорость: 127слов в секунду. Точность: -

Анализ

без  
без+PN;  
яшь  
яшь+Adj;яшь+N;  
чакта  
чак+N+LOC(ДА);  
чишмэ  
чиш+V+NEG(мА)+IMP\_SG();чишмэ+N;  
сулары  
су+N+PI(ЛАН)+POSS\_3(СЫ).

# tatcorp.antat.ru/text\_correction/1

Tat Corpora x

127.0.0.1:8000/corpus/text\_correction/1/

TatCorpora Тамгалау Корпус Инструментлар Проект турында Тел алмаштыру tatcorp

## Туган тел корпусы

Татар гомумтел корпусы

**Сайттагы корпус буенча мәгълумат**

Сайттагы текстлар саны: \*\*\*

Сүз формалары саны: ~\*\*\*

Күп таркалышлы төрләр саны: ~\*\*\*

Күп таркалышлы сүзләр саны: ~\*\*\*

## 111.txt

Урманда агач уон үсә . Анда барма 234а . Урман ( карурман ) ! Жир йөзөндә татар дигән халык бар . һәм татар дигән шушы 23123ваы халык , татар дигән шушы сүз

Артка Бетерү Бетерү һәм киләсе җөмлөләргә күчү

# Технологии

- Python 3.3
- Фреймворк Django 1.7
- СУБД PostgreSQL

A decorative element consisting of two light blue squares, one above and one below the green bar, positioned on the right side of the slide.A horizontal green bar with a slight 3D effect, spanning most of the width of the slide near the top.

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*