Опыт автоматизированной классификации контекстов при морфологической разметке корпуса

Анна Голуб, 11А
Научный руководитель:
Дмитрий Владимирович Грановский
Место практики:
проект «Открытый корпус» (OpenCorpora)

Что такое корпус?

Корпус — база текстов.

Например:

- собрание текстов на одном языке и их переводов на другой язык

- собрание текстов на данном языке

Что такое разметка?

Разметка — это информация о лингвистических характеристиках различных составляющих текста.

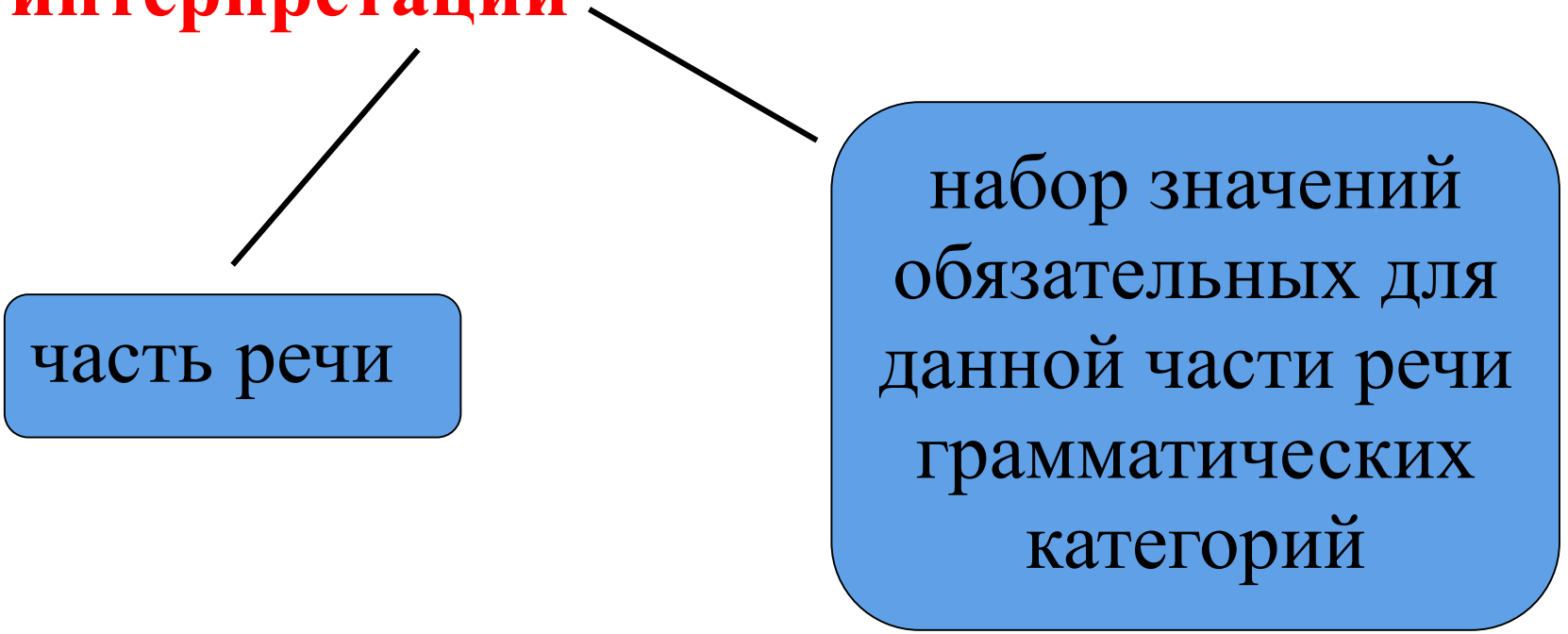
фонетическая - звуки

морфологическая - слова

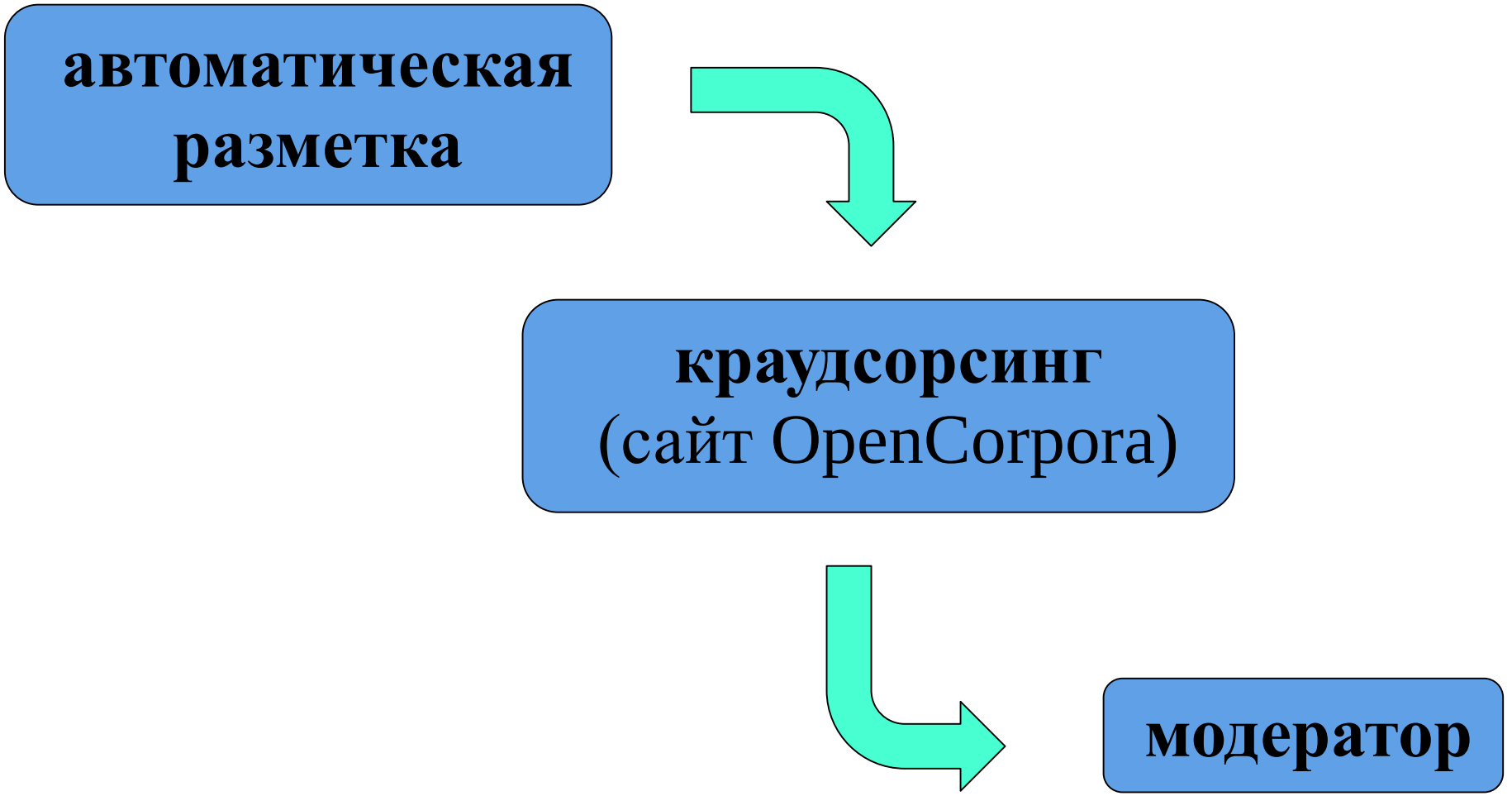
синтаксическая - предложения

Разметка OpenCorpora

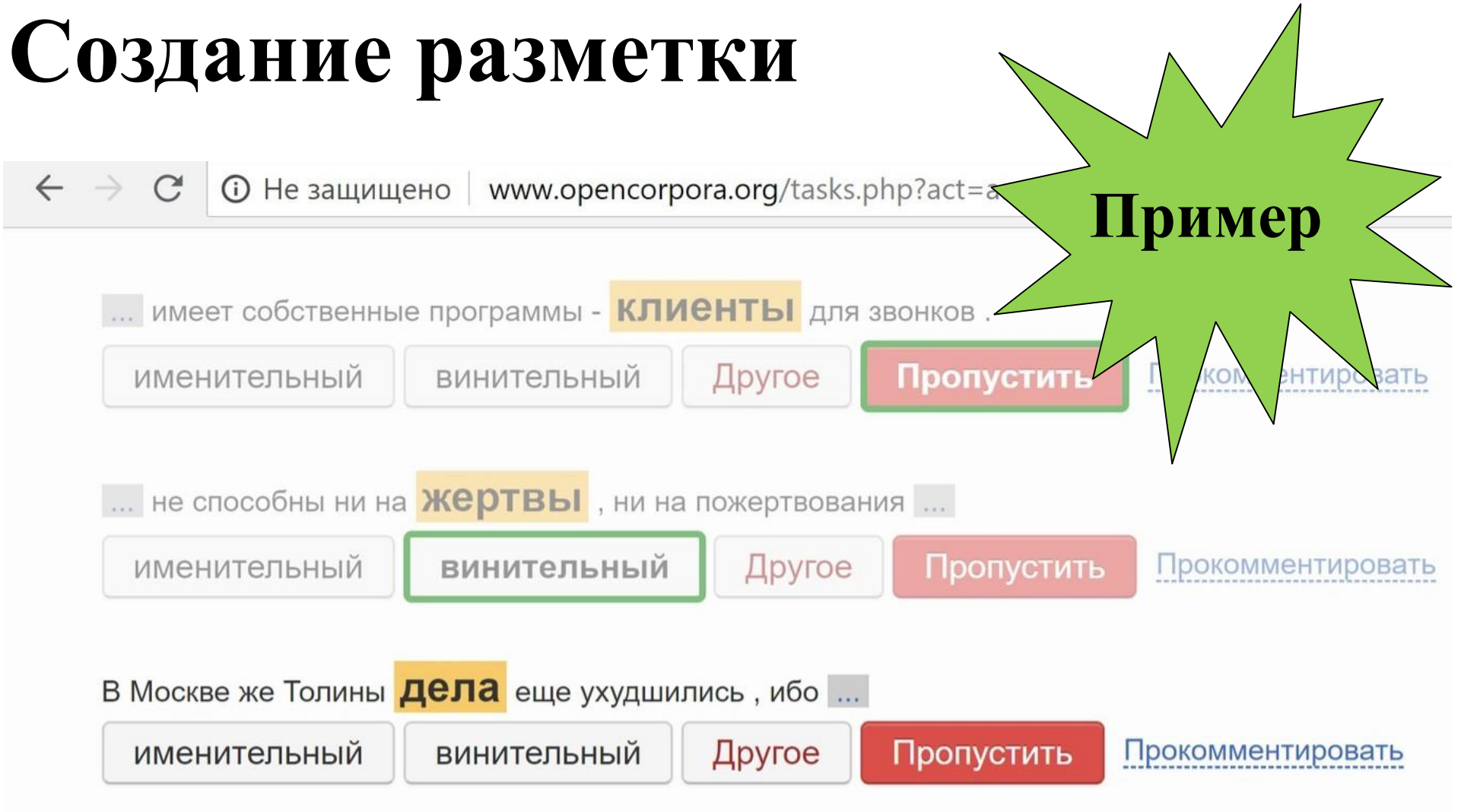
- **морфологическая**
- размечаемая единица – слово
- разметка каждого слова – одна/несколько **интерпретаций**



Создание разметки



Создание разметки



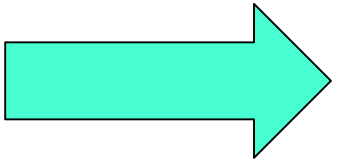
Создание разметки

<…> «Рождественскую песню в прозе: святочный **рассказ** с привидениями»

Постановка задачи

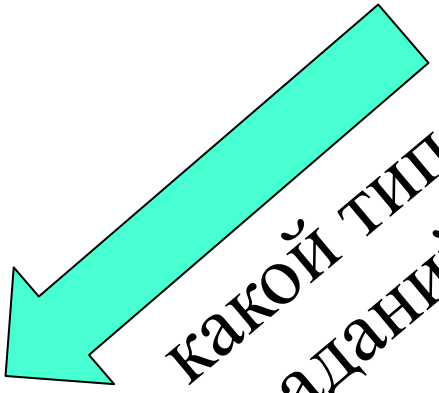
Фильтрация части заданий, попадающих к модератору.

путь решения



Выявление **критериев контекста**, дающих приемлемую вероятность неверного ответа.

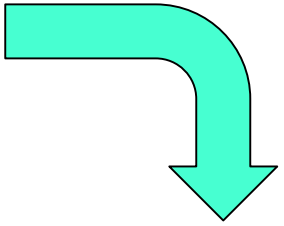
какой тип заданий



«именительный падеж»
vs
«винительный падеж»
для **существительных**

Методика

Оценить частоту случаев, где все размечавшие ошиблись в выборе ответа.



Разработать набор **критериев контекста**, по которым можно выделять такие случаи.

Результат

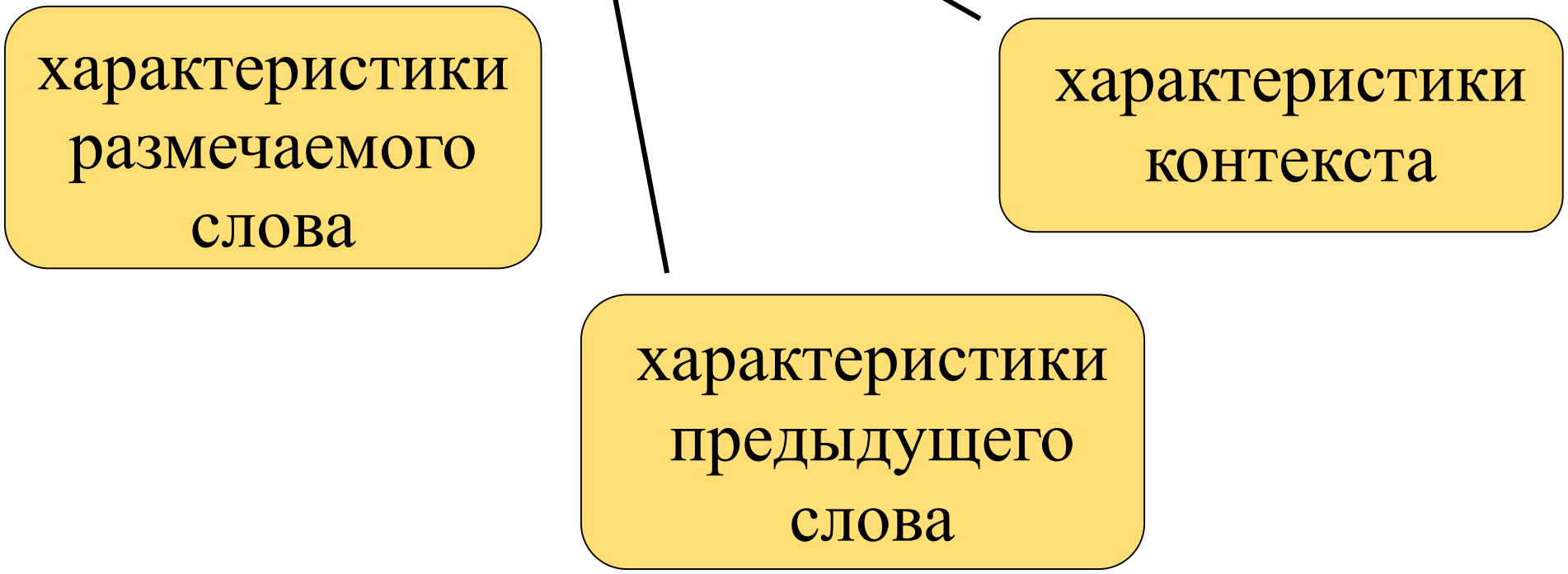
Оценить частоту случаев, где все размечавшие ошиблись в выборе ответа.

- Была рассмотрена выборка из 35199 примеров.
- Случаев, где все ответы размечавших – неверные, 0.2%. Это больше, чем пороговое значение 0.1%.
- Имеет смысл дальнейшее исследование.

Результат

Разработать набор **критериев контекста**, по которым можно выделять ошибкоопасные случаи.

Были разработаны **признаки контекста**



Результат

Разработать набор **критериев контекста**, по которым можно выделять ошибкоопасные случаи.

- Написан код, проверяющий выполнение признаков у каждого предложения.
- Результат его работы - файл, где в каждой строке записано предложение и бинарные коды признаков в определенном порядке.

Результат

Надо признать , обращение к современности выбило из рук писателя самый главный его козырь – [[знание]] скрупулезно описываемых реалий прошлого . 2 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Результат

Разработать набор **критериев контекста**, по которым можно выделять ошибкоопасные случаи.

- **Критерий контекста -** сочетание одного, двух или трех признаков.
- Выявлены критерии **наиболее ошибкоопасных** контекстов.

Результат

слово - аббревиатура

ФАС уже возбудила дело.

В США **сериал** шёл с большим успехом.

перед словом - аббревиатура

слово - в кавычках

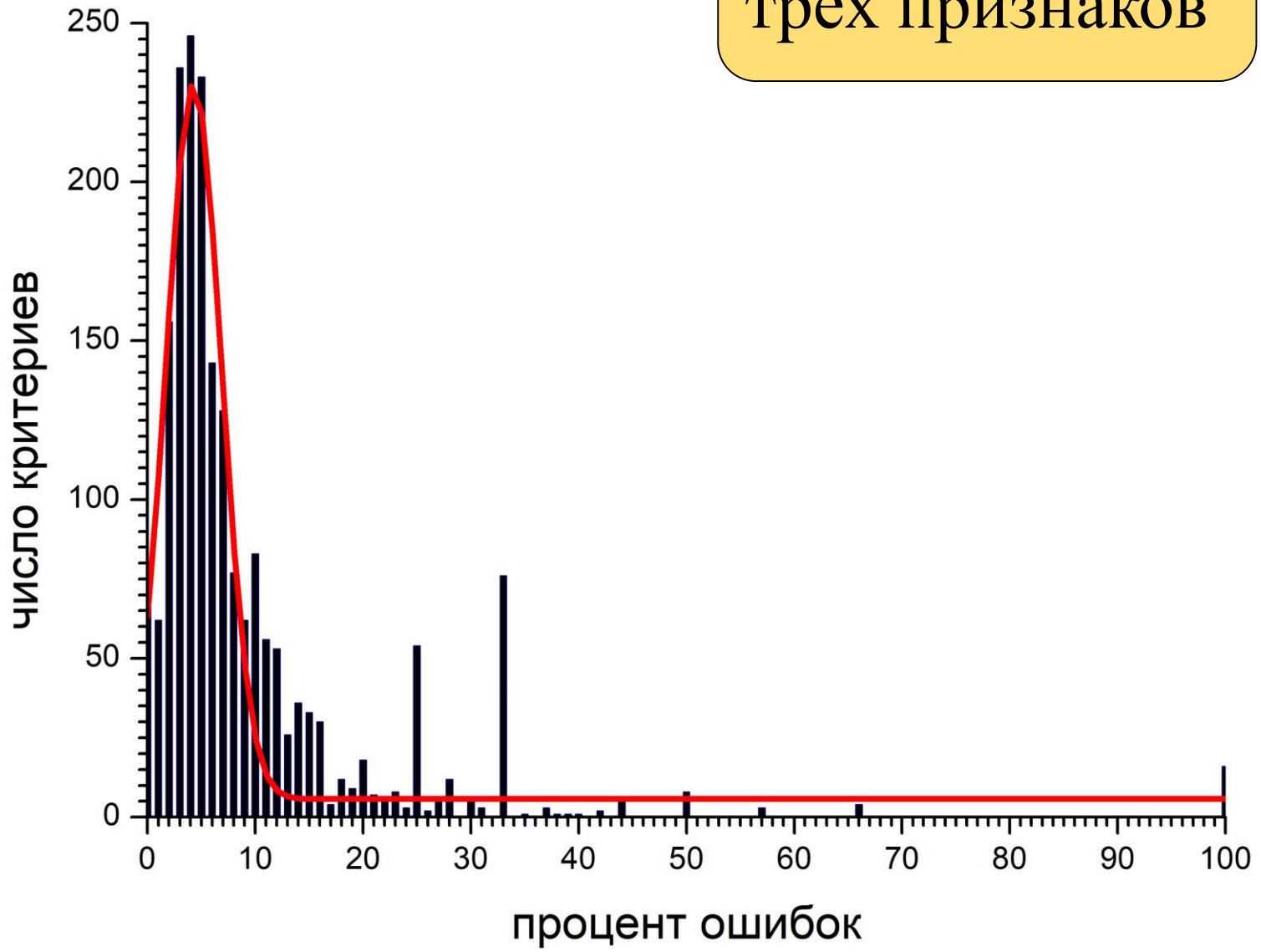
«**Ночь** в музее» сдвинули.

В итоге «Зенит» набрал 61 **очко**.

перед словом - число

Результат

критерии из трех признаков



Выводы

- Установлено, что существуют такие критерии контекста для заданий вида *"именительный падеж"* vs *"винительный падеж"* для существительных, что процент ошибок при выполнении соответствующего задания достаточно высок.
- Выявлены эти критерии.
- В дальнейшем эта информация может быть использована для автоматизации разметки *OpenCorpora*.


Спасибо
за внимание!