Crowdsourcing morphological annotation

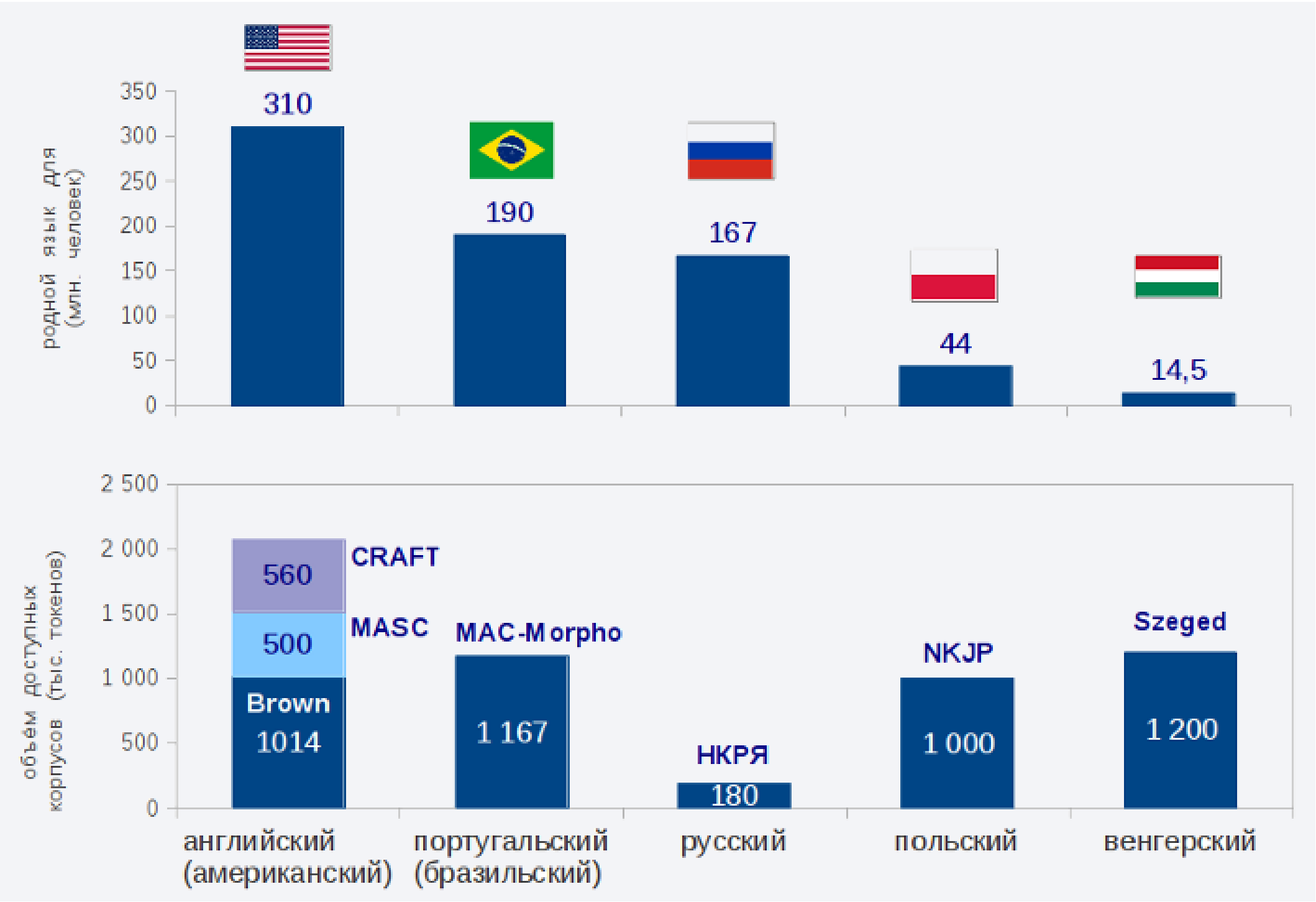
Bocharov V. V.,
Alexeeva S. V.,
Granovsky D. V.,
Protopopova E. V,
Stepanova M. E.,
Surikov A. V.

opencorpora.org
2.6.2013

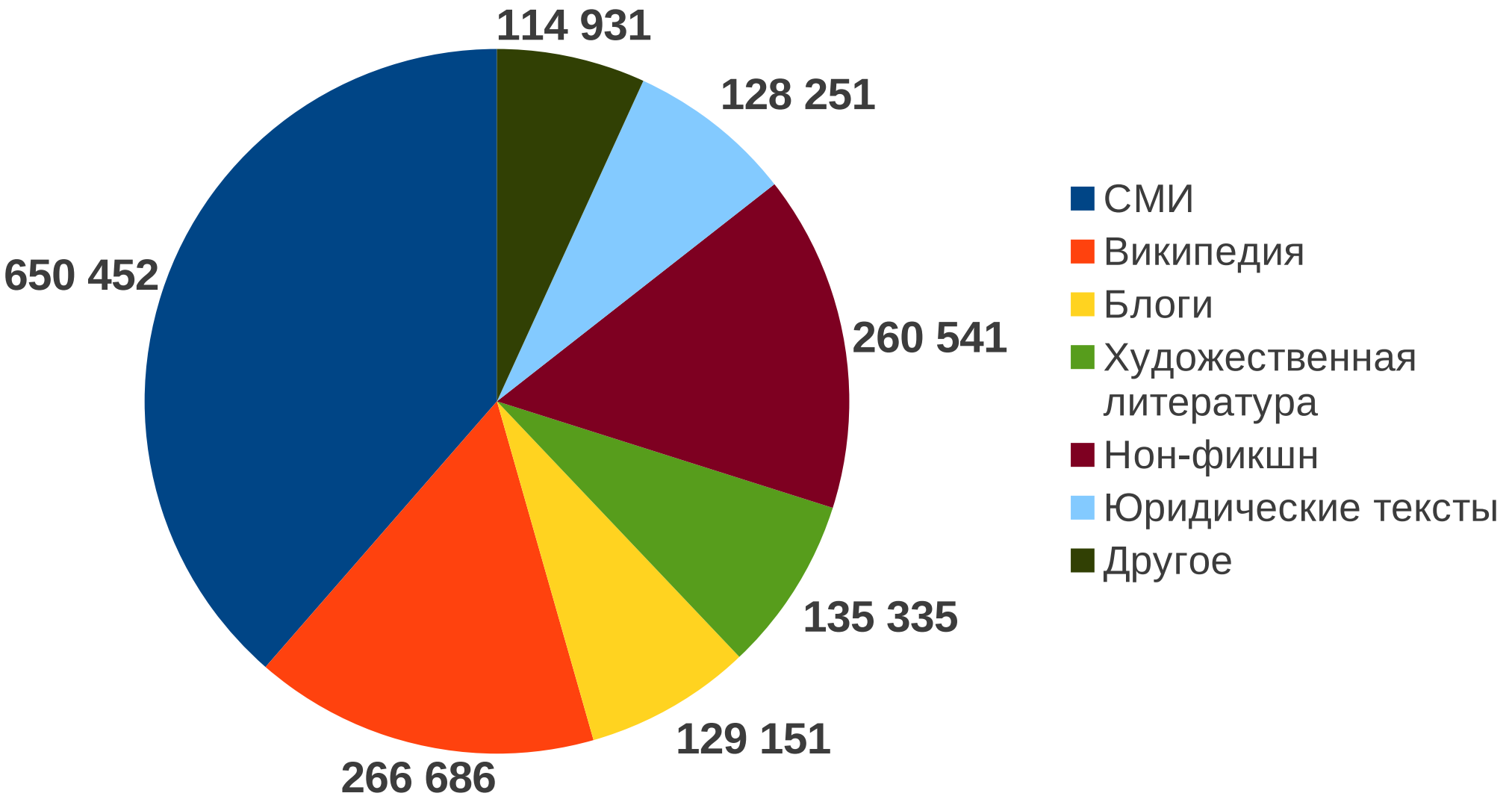
Цели проекта OpenCorpora

+ для автоматической обработки
+ высокая точность разметки
+ подробная разметка

- поиск примеров употребления
- большой размер корпуса


родной язык для (млн. человек)
350
300
250
200
150
100
50
0
310
190
167
44
14,5
объём доступных корпусов (тыс. токенов)
2 500
2 000
1 500
1 000
500
0
560
CRAFT
500
MASC
Brown
1014
MAC-Morpho
1 167
НКРЯ
180
NKJP
1 000
Szeged
1 200
английский (американский)
португальский (бразильский)
русский
польский
венгерский

Состав корпуса



Разметка

- 100% метаразметка
- 100% сегментация
- морфология
- синтаксис ?
- лексическая семантика ?

Мама	мыла	раму
v мама x СУЩ, од, жр, ед, им	v мыло x СУЩ, неод, ср, ед, рд	v рам x СУЩ, неод, мр, гео, ед, дт
	v мыло x СУЩ, неод, ср, мн, им	v рама x СУЩ, неод, жр, ед, вн
	v мыло x СУЩ, неод, ср, мн, вн	
	v мыть x ГЛ, несов, перех, жр, ед, прош, изъяв	

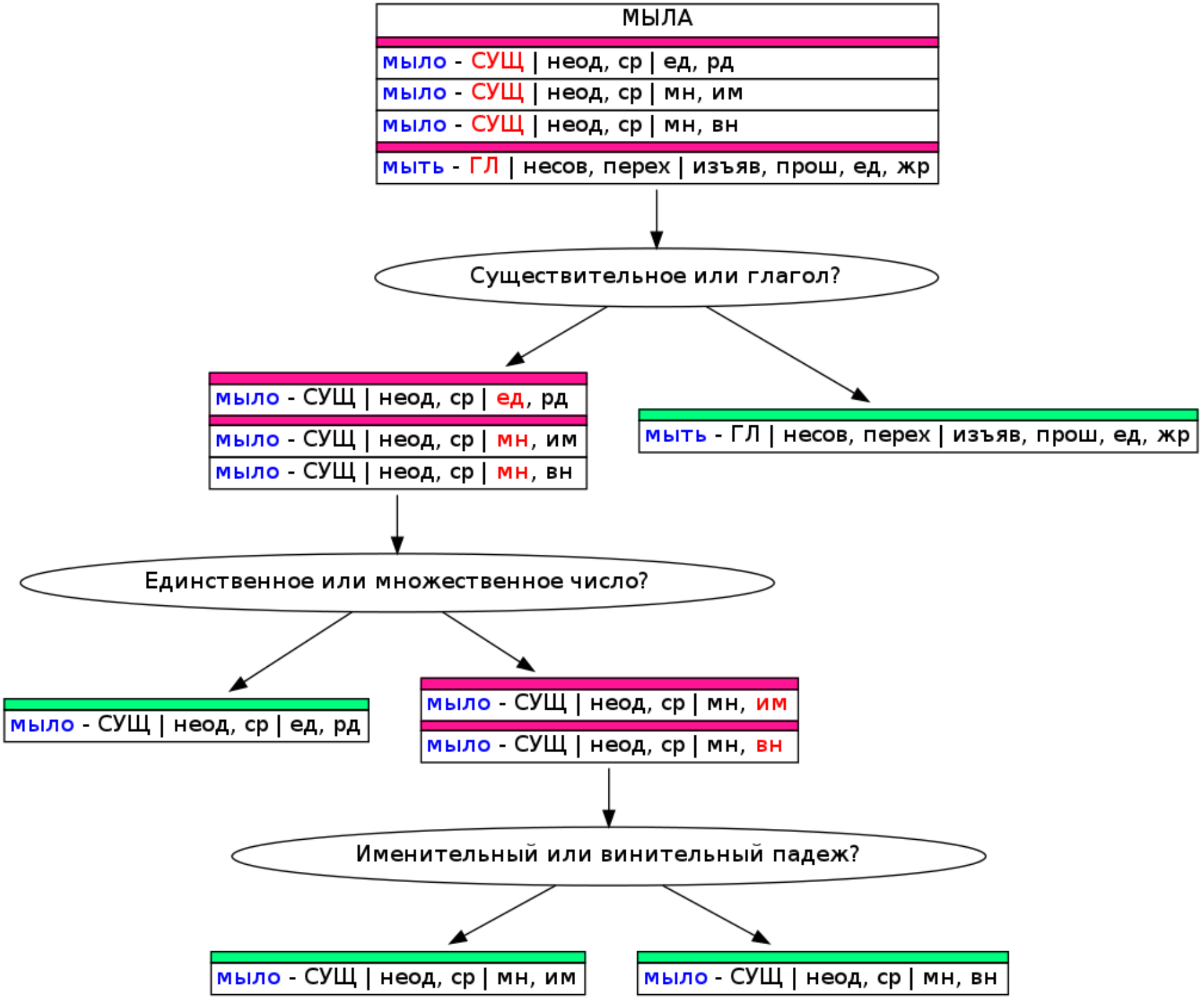
Мама	мыла	раму
v мама x СУЩ, од, жр, ед, им	v мыть x ГЛ, несов, перех, жр, ед, прош, изъяв	v рама x СУЩ, неод, жр, ед, вн

Морфологическая разметка

- простые однотипные задания
- разметка несколькими участниками
- модерация
- автоматический контроль

Жизненный цикл задания

- Поиск заданий по условию
  - пулы заданий (по 200 штук)
- Выполнение задания
- Модерация ответов
- Разметка корпуса
  - пул → корпус
- Архивирование пула


МЫЛА
мыло - СУЩ | неод, ср | ед, рд
мыло - СУЩ | неод, ср | мн, им
мыло - СУЩ | неод, ср | мн, вн
мыть - ГЛ | несов, перех | изъяв, прош, ед, жр
Существительное или глагол?
мыло - СУЩ | неод, ср | ед, рд
мыло - СУЩ | неод, ср | мн, им
мыло - СУЩ | неод, ср | мн, вн
мыть - ГЛ | несов, перех | изъяв, прош, ед, жр
Единственное или множественное число?
мыло - СУЩ | неод, ср | ед, рд
мыло - СУЩ | неод, ср | мн, им
мыло - СУЩ | неод, ср | мн, вн
Именительный или винительный падеж?
мыло - СУЩ | неод, ср | мн, им
мыло - СУЩ | неод, ср | мн, вн

Существительное: единственное / множественное число	инструкция	Взять на разметку
Существительное / Предлог	инструкция	Взять на разметку
Союз / Междометие / Частица	инструкция	Взять на разметку
Существительное, мн. ч.: именительный / винительный	инструкция	Взять на разметку
Существительное, ед. ч.: родительный / винительный	инструкция	Взять на разметку
Существительное, ед. ч.: именительный / винительный	инструкция	Взять на разметку
Существительное, мн. ч.: родительный / винительный	инструкция	Взять на разметку
Прилагательное: мужской / средний род		Взять на разметку
Прилагательное: единственное / множественное число		Взять на разметку
Инфинитив / Существительное		Взять на разметку
Существительное, ед. ч.: дательный / предложный		Взять на разметку
Существительное / Деепричастие		Взять на разметку
Причастие: мужской / средний род		Взять на разметку
Глагол / Числительное		Взять на разметку

Спасибо, что помогаете нам. Не торопитесь, будьте внимательны. Если вы не уверены, пропускайте пример.

... а какие-то внутренние психологические **пересечения** есть и с ним ...

единственное число | множественное число | Другое | Пропустить | Прокомментировать

... 3,6 см вблизи апоцентра **кольца** .

единственное число | множественное число | Другое | Пропустить | Прокомментировать

... сотрудничество с Пакистаном в **области** борьбы с терроризмом помогло ...

единственное число | множественное число | Другое | Пропустить | Прокомментировать

... Канады темы , популярность **группы** вышла далеко за пределы ...

единственное число | множественное число | Другое | Пропустить | Прокомментировать

Компания **ООО** « Нефтестрой » .

единственное число | множественное число | Другое | Пропустить | Прокомментировать

Хочу ещё примеров! | Спасибо, достаточно

id	Текст + варианты разбора	1	2	3	Модератор
7349	... в свете первой удачной **[картины]** – закономерность . картина, NOUN, inan, femn, sing, gent картина, NOUN, inan, femn, plur, nomn картина, NOUN, inan, femn, plur, accs	NOUN & sing	NOUN & sing	NOUN & sing	NOUN & sing
7350	... Питера FM » от **[Оксаны]** Бычковой хотели истории в ... оксана, NOUN, anim, femn, Name, sing, gent оксана, NOUN, anim, femn, Name, plur, nomn	NOUN & sing	NOUN & sing	NOUN & sing	NOUN & sing
7351	... от Оксаны Бычковой хотели **[истории]** в этом же ключе ... история, NOUN, inan, femn, sing, gent история, NOUN, inan, femn, sing, datv история, NOUN, inan, femn, sing, loct история, NOUN, inan, femn, plur, nomn история, NOUN, inan, femn, plur, accs	NOUN & sing	NOUN & sing	NOUN & plur	NOUN & sing

Модерация

Ожидаемая точность:

10% ошибок * 3 участника

= 0.1% случаев, когда трое ошиблись

т. е. 1 ошибка / 1000 примеров

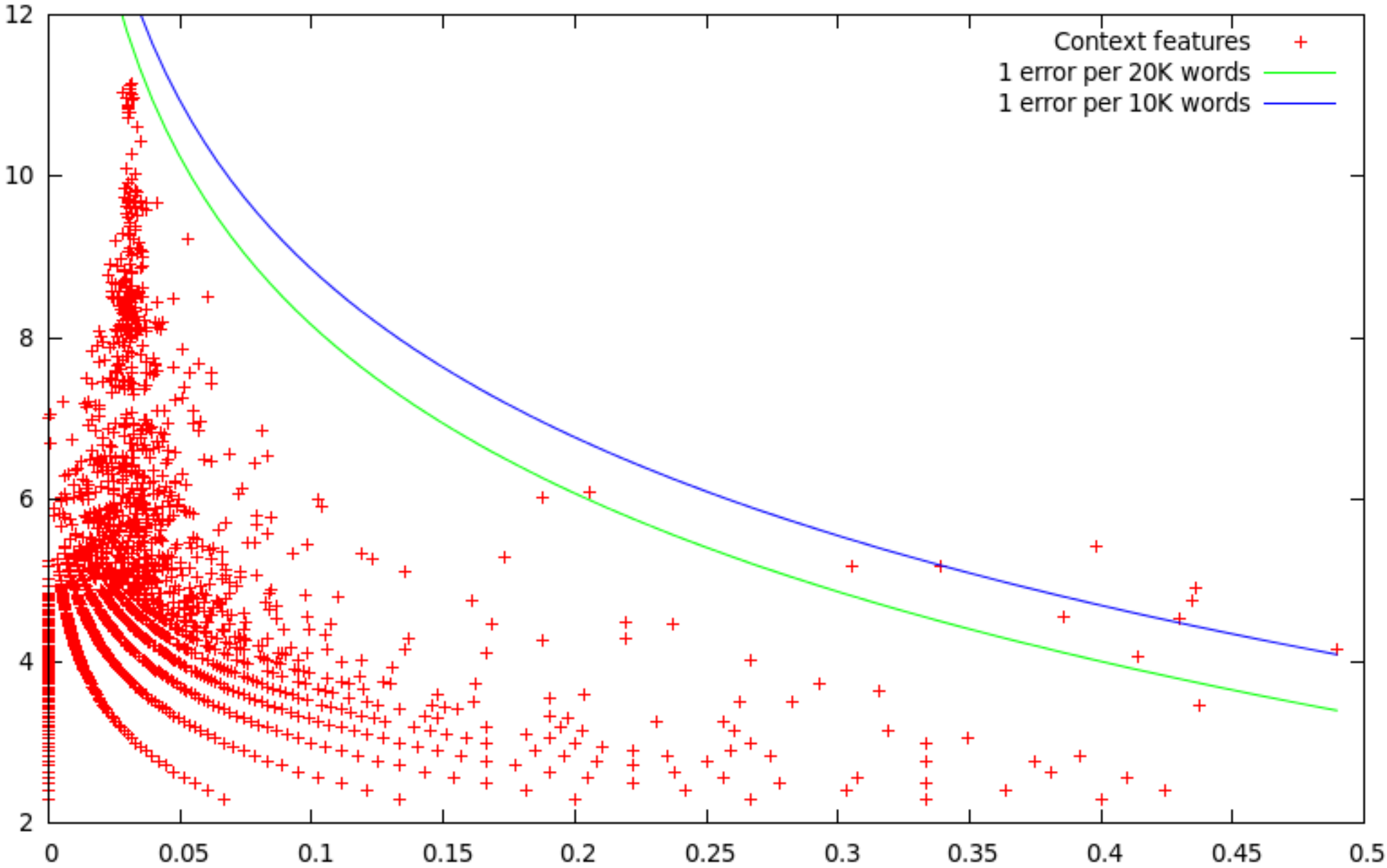
Реальная точность:

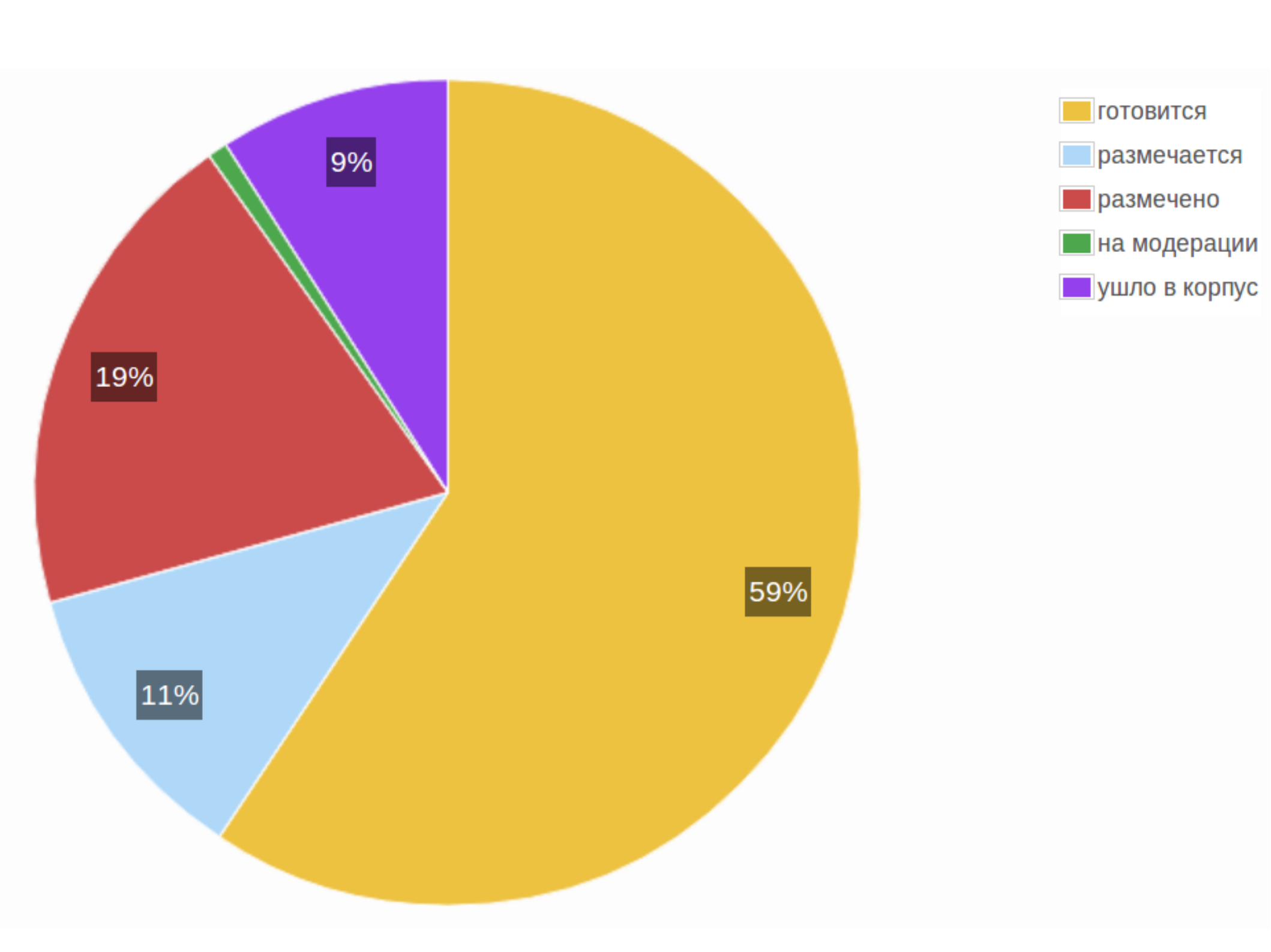
0.5 — 10% ошибок

2 % случаев, когда трое ошиблись

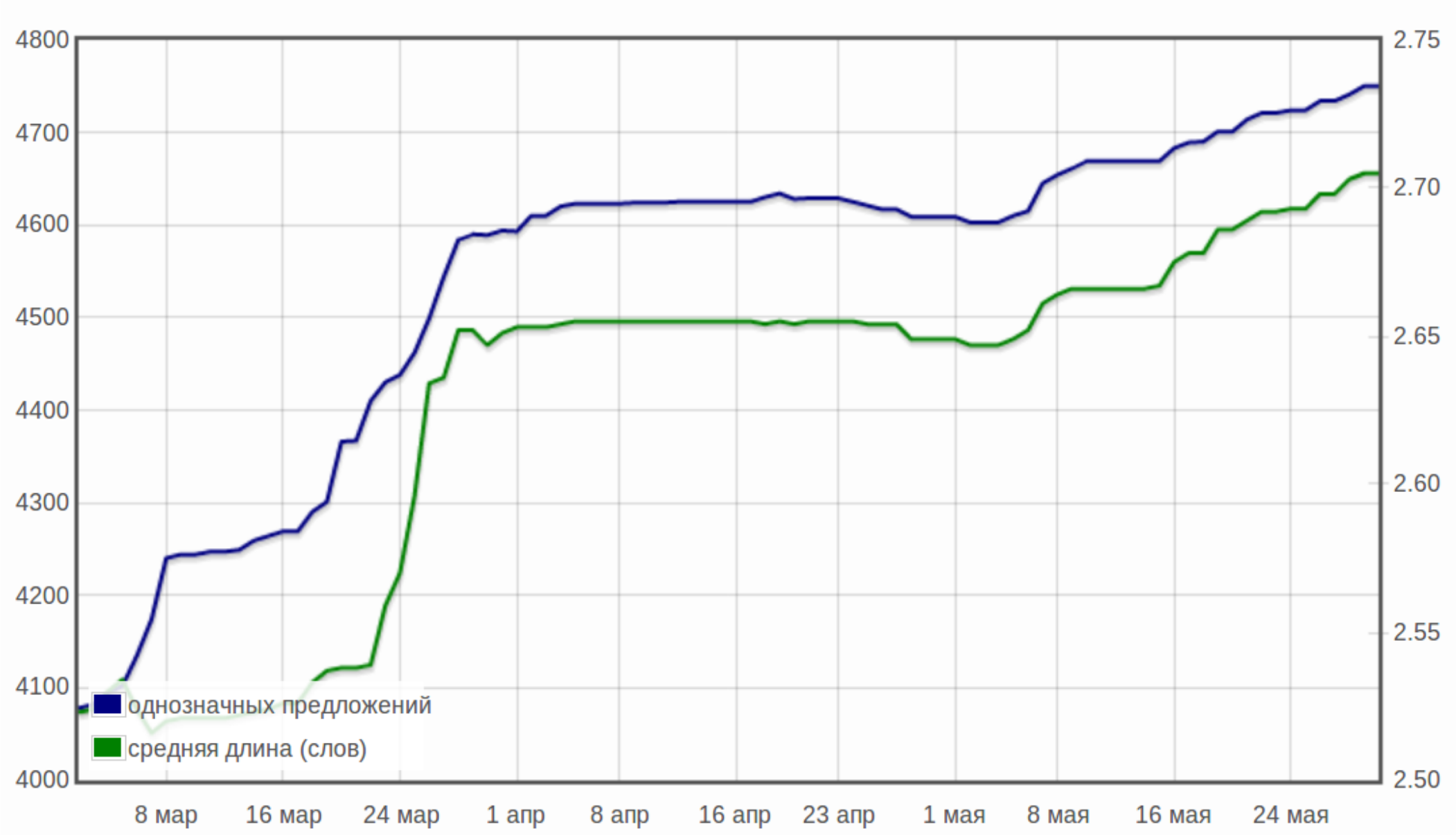
Модерация

Context feature	Position	Total samples	Samples with disagreement	Samples without disagreement	Disagreement rate	Expected error probability
word=четыре	-1	64	47	17	73,44%	48,96%
word=две	-1	136	89	47	65,44%	43,63%
word=три	-1	115	75	40	65,22%	43,48%
word=два	-1	93	60	33	64,52%	43,01%
word=две	-2	58	36	22	62,07%	41,38%
word=одна	4	13	8	5	61,54%	41,03%
word=две	0	226	135	91	59,73%	39,82%
word=копейки	0	17	10	7	58,82%	39,22%
word=четыре	-	95	55	40	57,89%	38,60%

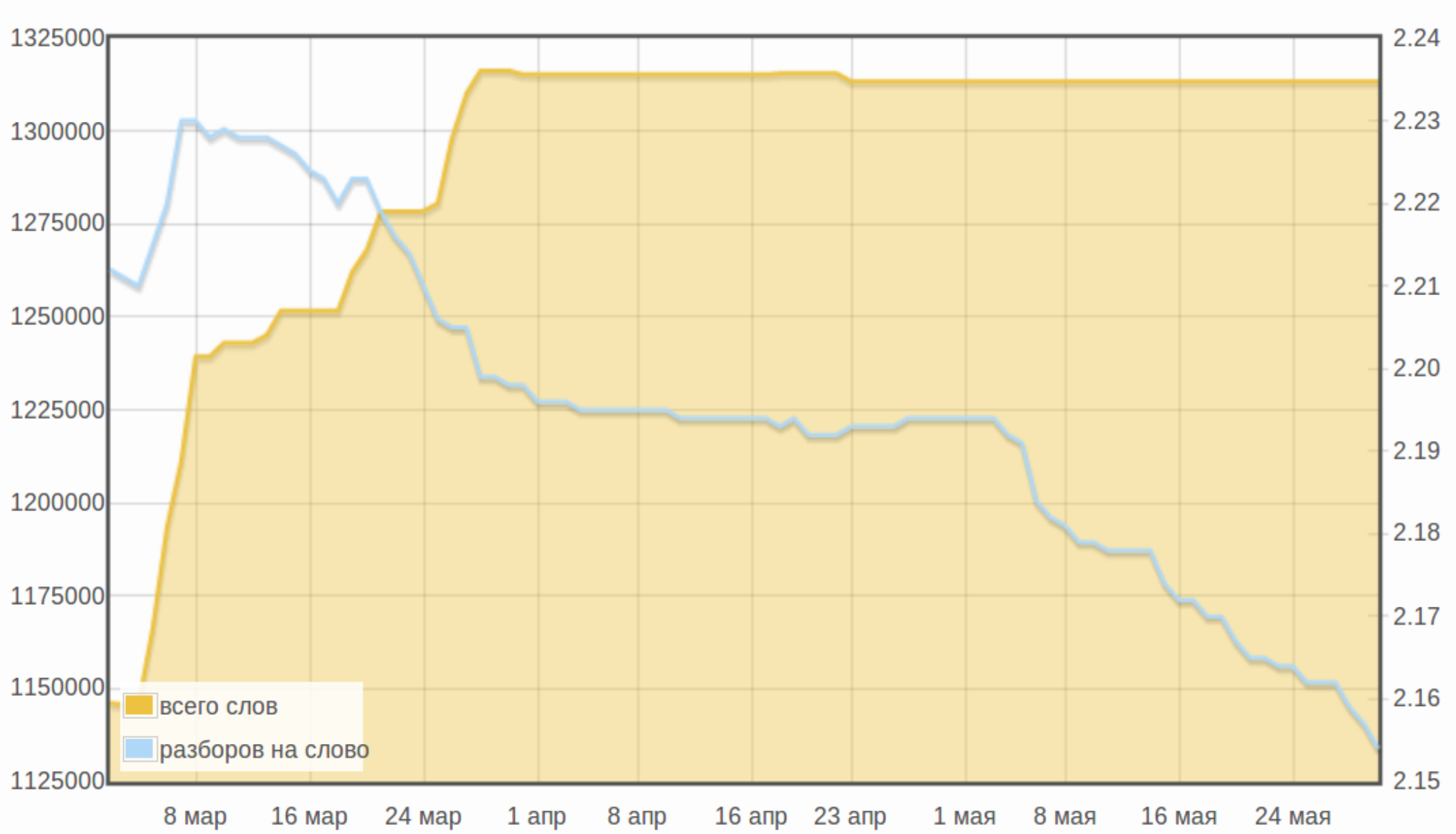

Context features
1 error per 20K words
1 error per 10K words
12
10
8
6
4
2
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5


9%
19%
59%
11%
готовится
размечается
размечено
на модерации
ушло в корпус

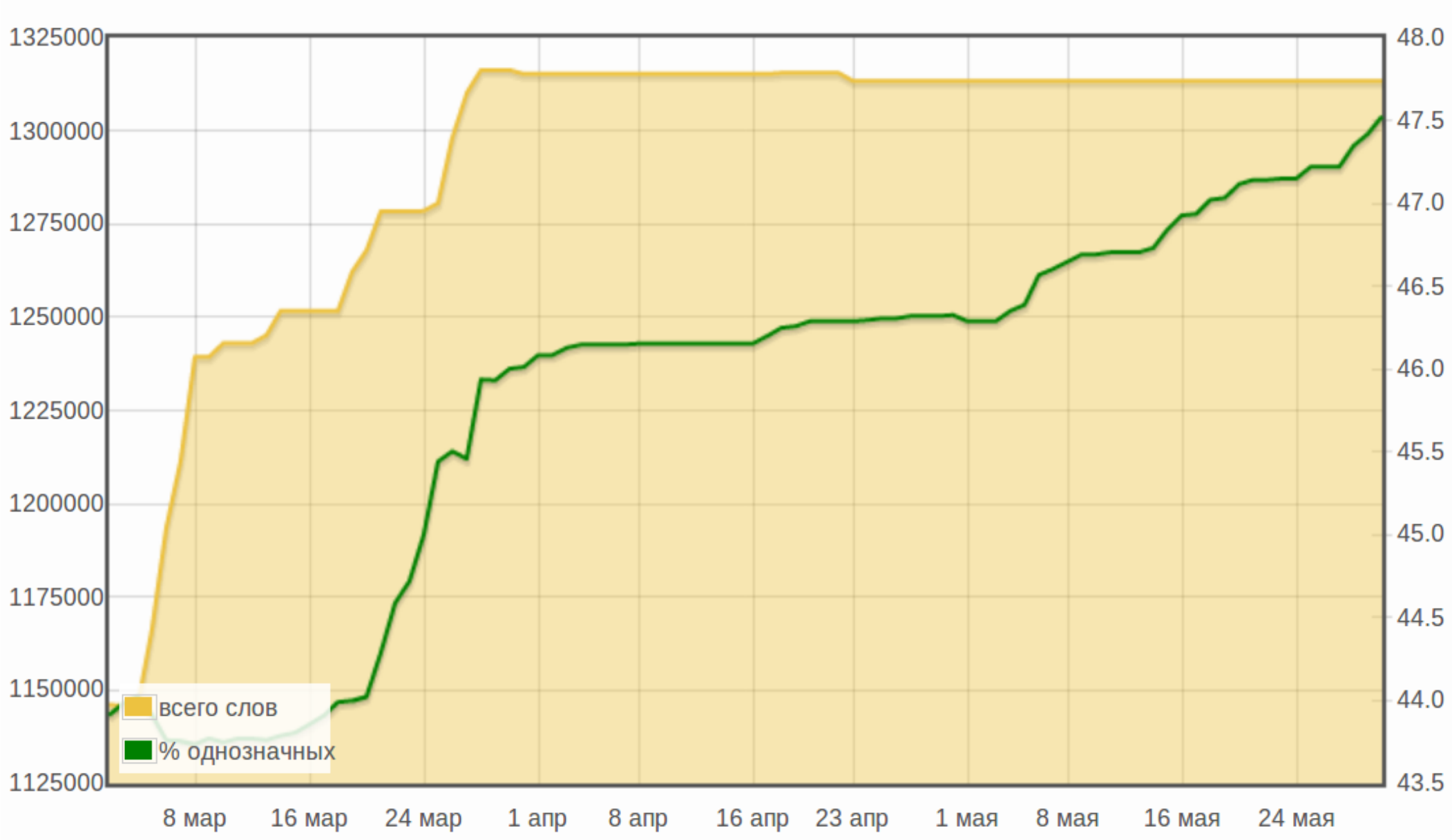
Однозначные разборы



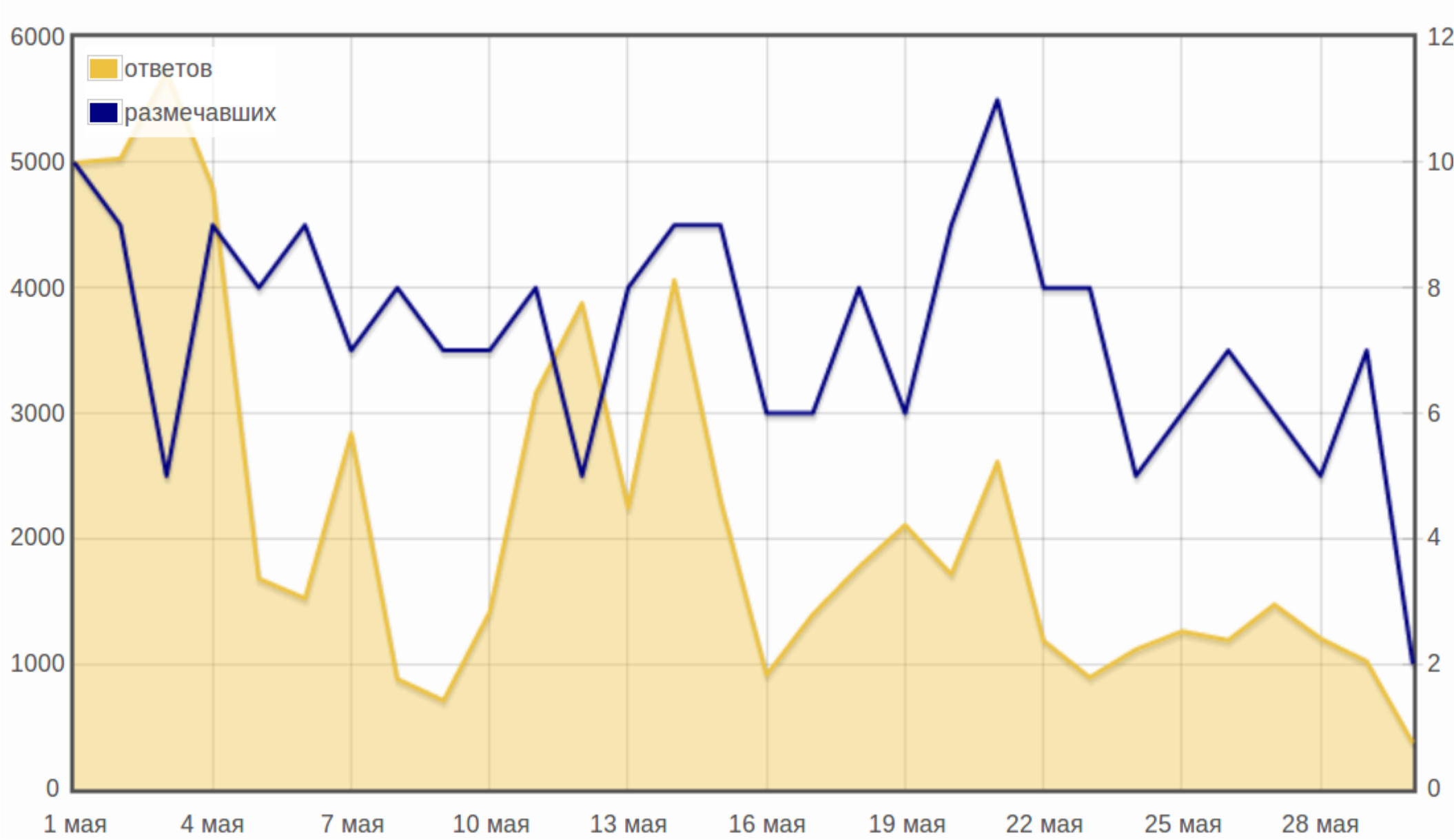
Количество разборов на слово



Количество однозначных слов

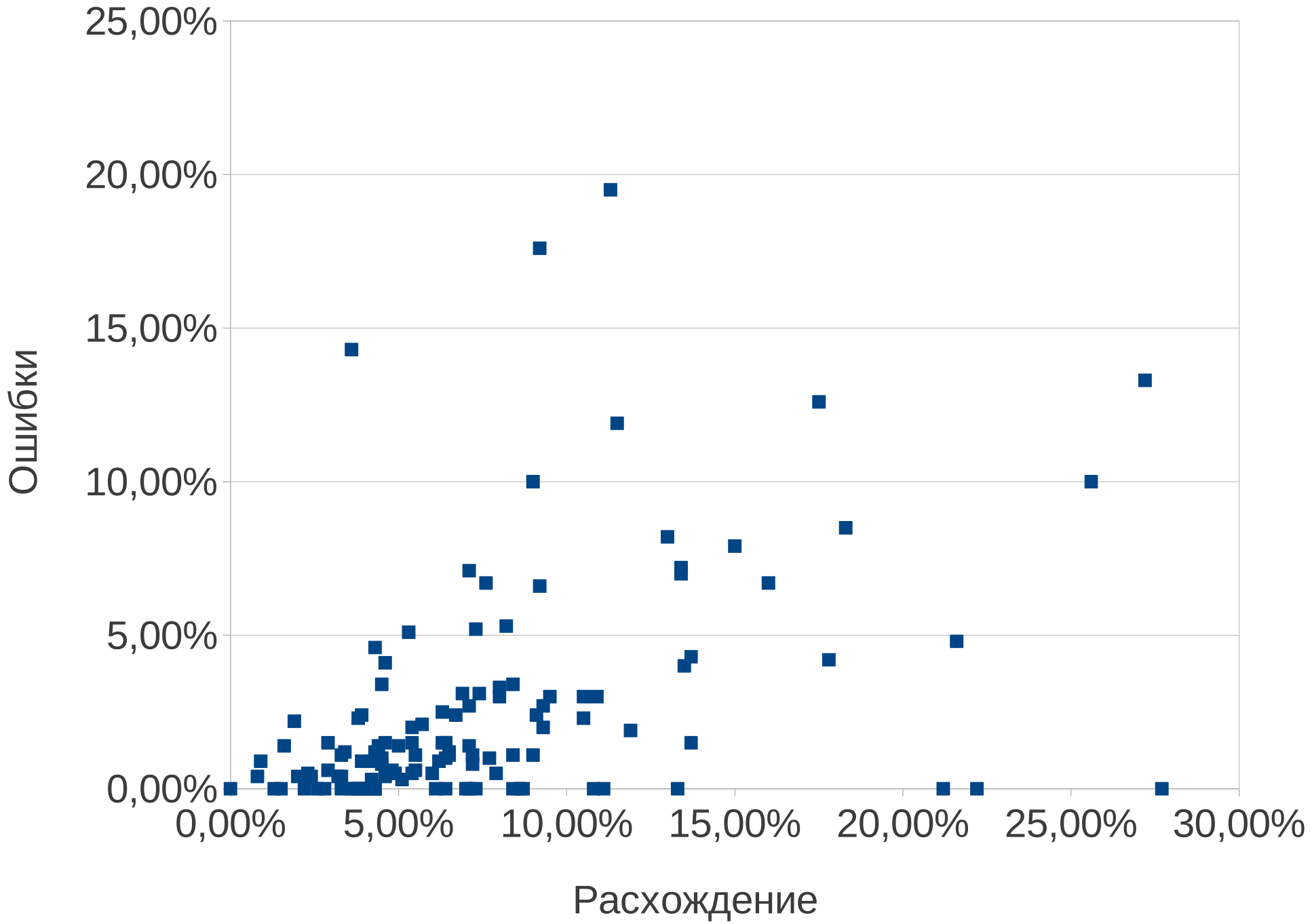


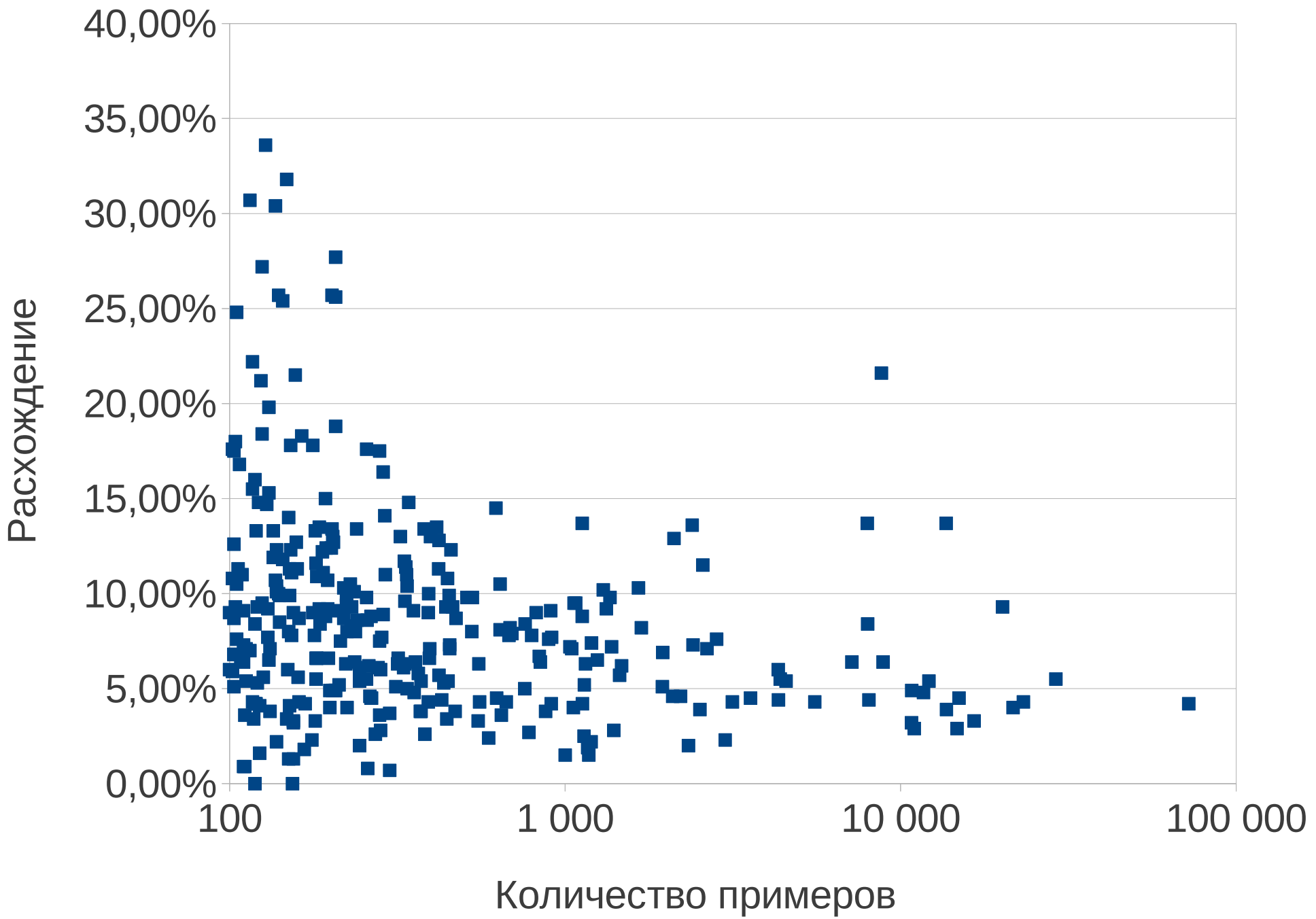
Активность участников



Участники

#	Участник	Рейтинг	Ответов	В завершённых пулах		В проверенных пулах		Последняя активность
				Размечено	% расхождений	Размечено	% ошибок	
1	Lvova	545499	359367	232205	4.1%	62783	0.7%	сегодня в 09:01
2	Nofenigma	174551	109663	100146	4.2%	30686	1.8%	27.05.13
3	Мяу	114471	103877	93993	2.2%	42936	0.6%	08.03.13
4	Rave	87745	83522	64597	3.5%	25730	0.6%	19.05.13
5	quorax	75832	38757	34580	4.3%	6441	0.6%	19.02.13


25,00%
20,00%
15,00%
10,00%
5,00%
0,00%
Ошибки
0,00%
5,00%
10,00%
15,00%
20,00%
25,00%
30,00%
Расхождение


40,00%
35,00%
30,00%
25,00%
20,00%
15,00%
10,00%
5,00%
0,00%
Расхождение
100
1 000
10 000
100 000
Количество примеров

Данные

- Корпус с частично снятой омонимией
- Подкорпус с полностью снятой омонимией
- Ответы на задания (-)
- Хронология разметки (- -)
- Словарь

Что дальше?

- Разметить какие-то задания
- Рассказать коллегам
- Разрешить использование своих текстов на условиях Creative Commons

- twitter.com/opencorpora
- vk.com/opencorpora
- opencorpora@opencorpora.org
- http://opencorpora.org/?page=about

Спасибо!