Русский классический стих в свете корпусных данных и глубокого обучения

Борис Орехов

НИУ ВШЭ, ИРЛИ РАН

nevmenandr@gmail.com

19 сентября 2023

Содержание

Содержание

Особая дисциплина

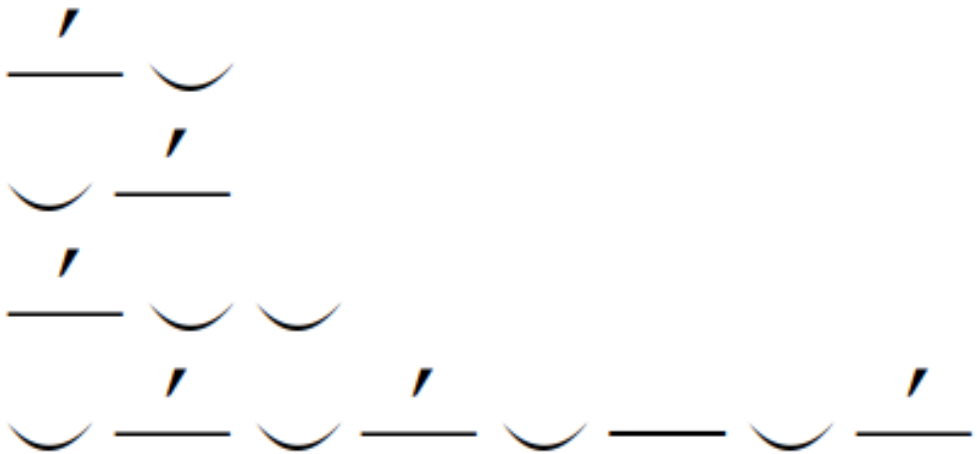
- Не всякое изучение стихов = стиховедение.
- Стиховедение (verse studies) — особая дисциплина, локализуется в области специфических особенностей стихотворной формы.
- В силу высокой степени формализации материала имеет давние квантитативные традиции.
- Популярный курс на Постнауке: https://postnauka.org/themes/orehovb

Содержание

Силлаботоника

- Хорошо исследованный материал.
- Классик социологии Харви Сакс считал, что бывает «естественная анализируемость» социальных действий.
- Силлабо-тоника обладает «естественной» формализуемостью.
- Формализуемость поддается квантификации.
- Деление обусловлено исторически.

Составные элементы силлаботоники



Выработанное месторождение

«Естественные» формальные параметры силлаботоники так подробно обсчитаны, что М. Л. Гаспаров в свое время высказался радикально: наука о стихе разработана настолько хорошо, что революций там не предвидится. «Методика исследования уже выработана, и здесь требуются только время и способные аспиранты».

Начало 2000-х годов.

Данные о силлаботонике

Таблица 3

Метрический репертуар поэзии XVIII—XX вв. (лирика): число текстов

Размер	1740—1800	1800—1830	1830—1880	1880—1900	1890—1935	1936—1957	1958—1980
Я 3-ст.	40	33	46	8	61	92	102
4-ст.	242	622	828	259	530	499	625
5-ст.	2	116	296	68	280	548	631
6-ст.	326	136	351	162	107	21	28
рз.	29	133	238	67	105	150	98
вол.	546	418	175	52	139	82	43
проч.	—	10	11	4	8	3	6
Всего ямбов	1185	1468	1945	620	1239	1385	1533
Х 3-ст.	2	15	113	6	25	45	16
4-ст.	108	181	521	111	227	236	215
5-ст.	1	3	46	30	170	334	253
6 цез.	—	—	46	16	19	8	5
6 б/ц	—	9	24	5	26	15	9
рз.	15	24	113	28	59	98	39
проч.	4	1	16	5	67	12	21
Всего хореев	130	233	879	201	593	748	558

После Гаспарова

- Появился поэтический корпус в составе НКРЯ.
- Стали доступнее количественные методы исследования и обработки данных.

Содержание

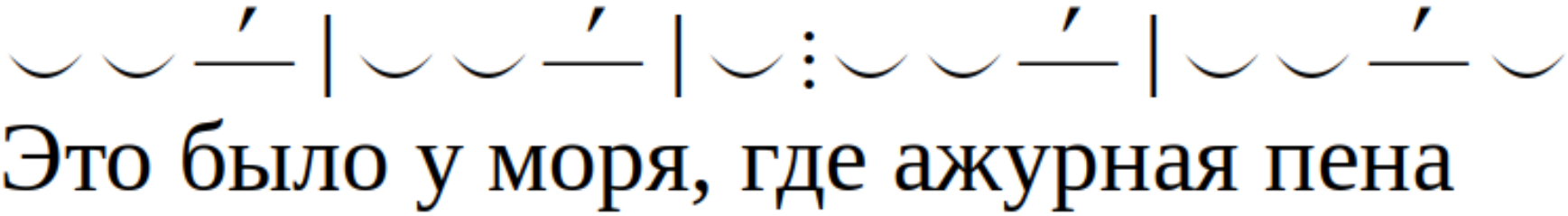
Докорпусный подход

§ 1. Мы видели, что у нас имеется пять видов стоп. Стихотворная строчка может вмещать практически от двух до восьми стоп; таким образом, по числу стоп, каждый размер может быть семи видов. Значит, мы получаем (5×7) тридцать пять размеров. Каждый размер, по окончанию, может быть семи видов (мужская строчка, женская, дактилическая, гипердактилическая с тремя безударными слогами, с четырьмя, с пятью, с шестью; более длинные окончания на практике неосуществимы, хотя можно подобрать несколько слов с еще более длинным «хвостом», напр.: **вЫкристаллизовавшимися,**—где после ударения идет восемь безударных слогов. Значит, с учетом окончаний, мы получаем (35×7)=245 размеров. Далее, приблизительно половина этих размеров имеет обязательную большую цезуру, а при ее наличии первое полустишие может быть полномерным, усеченным, наращенным на один, два, три—редко более—слога;

Шенгели Г. Техника стиха. М., 1940. С. 23.

Цезурные эффекты

Цезура (в применении к русскому классическому стиху) — это регулярный словораздел.



Возможны:

- Цезурное наращение
- Цезурное усечение

Стих продолжает оставаться регулярным силлабо-тоническим.

Подсчеты Шенгели и корпусные данные 1/2

Шенгели

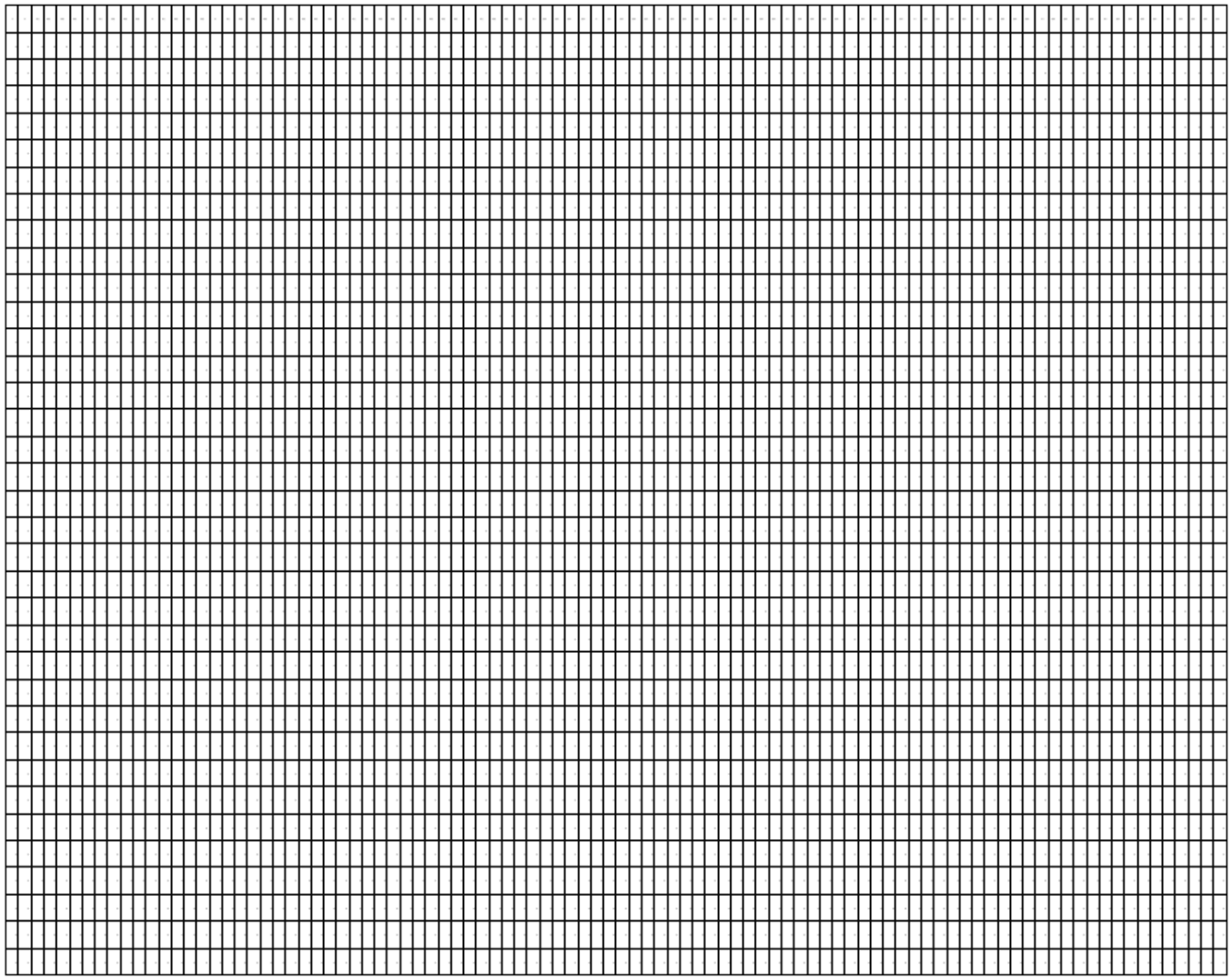
- Стопность: от 2 до 7: $5 * 7 = 35$
- 7 видов клаузул: м, ж, д, г3, г4, г5, г6: $35 * 7 = 245$
- 4 вида цезурных эффектов: цу, цн1, цн2, цн3: $120 * 4 = 480$
- Всего: $480 + 125 = 605$

Подсчеты Шенгели и корпусные данные 1/2

Корпус

- Стопность: от 1 до 20.
- Диапазон длины гипердактилической клаузулы подтвердился (в абсолютных числах):
  - г3: 4373,
  - г4: 349,
  - г5: 41,
  - г6: 17

Матрица потенциальности силлабо-тоники



Матрица потенциальности силлабо-тоники

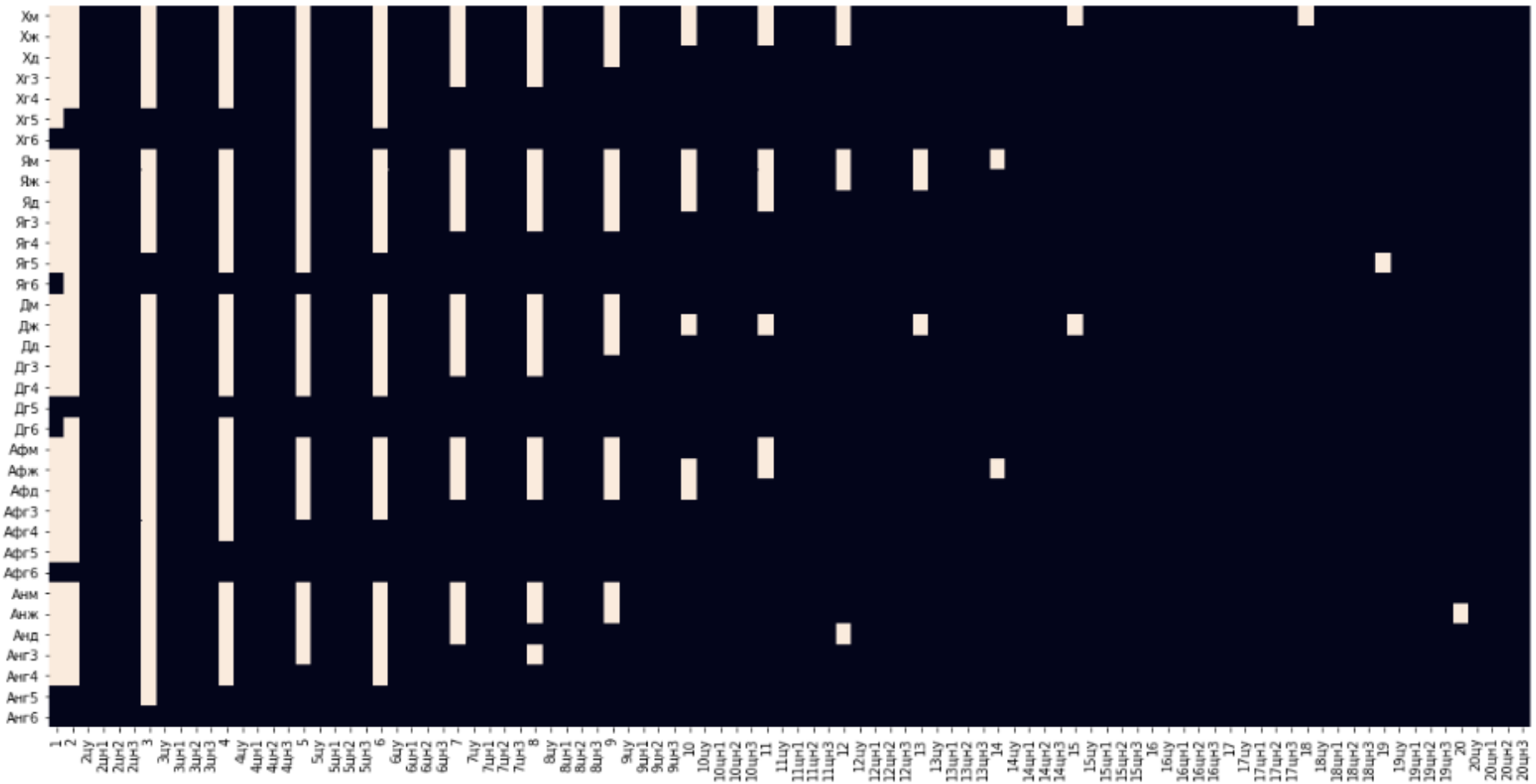
Колонки: стопность + цезурные эффекты: 1, 2, 2цу, 2цн1, ... 12цн3

Строки: метр + клаузула: Хм, Хж, Хд, Хг3, Хг4, ... Ям, Яж, Яд, Яг3, Яг4, Яг5, Яг6 ...

35 строк $\times$ 96 колонок $=$ 3360

Реальные размеры в корпусе

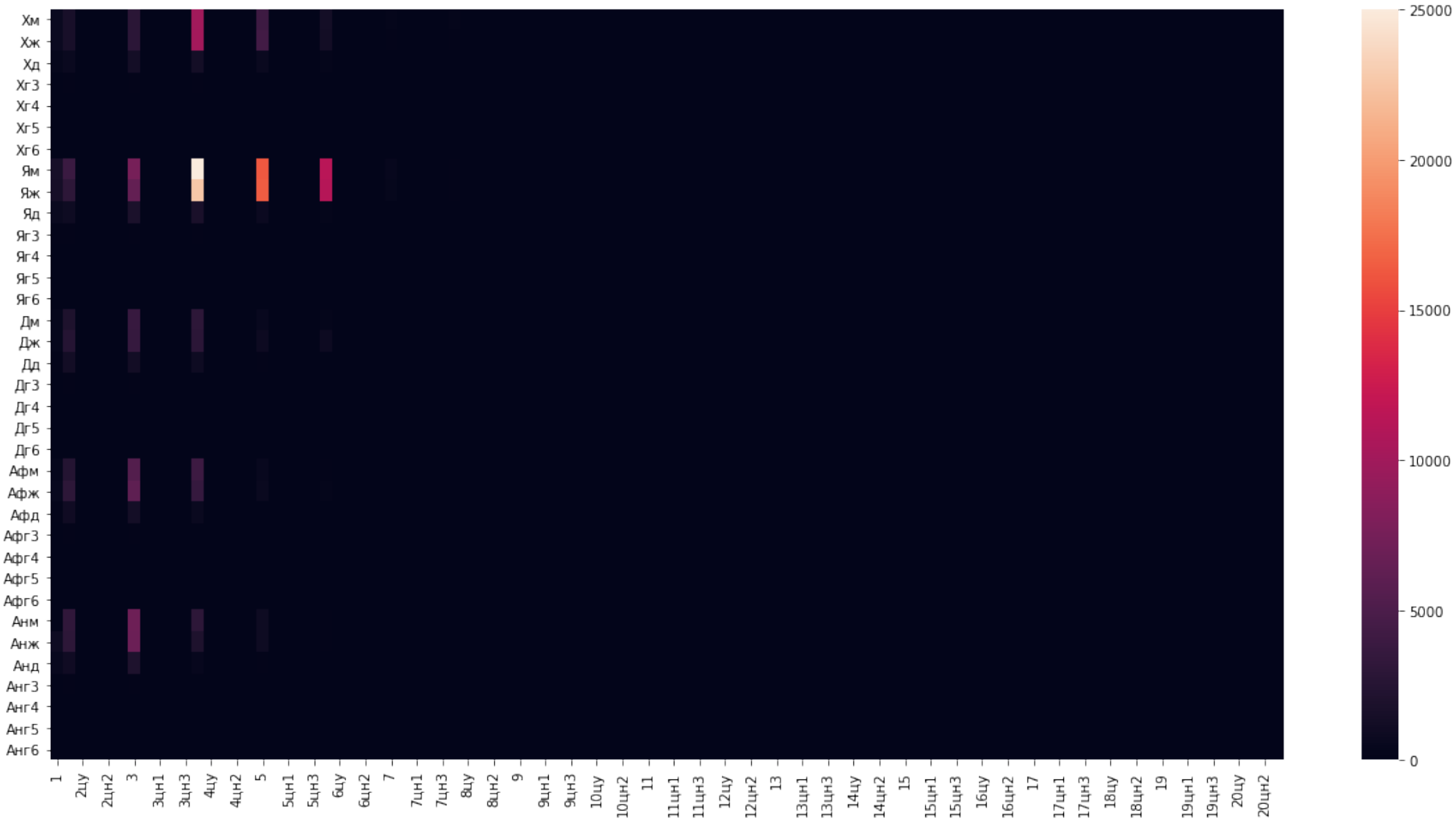
Без цезурных эффектов



«Бинарная» тепловая карта: 248 размеров

Реальные размеры в корпусе

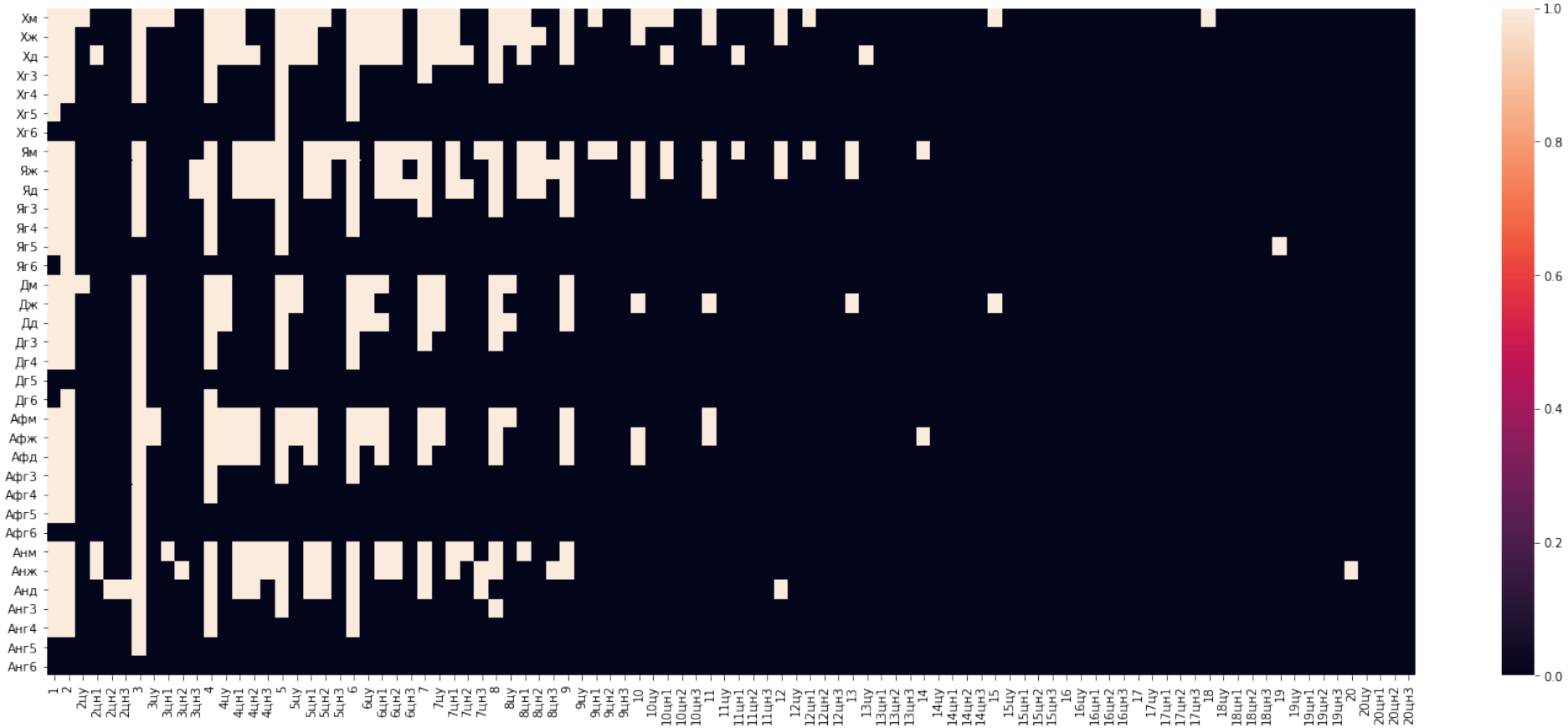
Без цезурных эффектов



Настоящая тепловая карта

Реальные размеры в корпусе

С цезурными эффектами: +362



«Бинарная» тепловая карта

Корпус и Шенгели

- Уточнение: не может быть цезурных усечений у Я и Ан. $3360 - 266 = 3094$
- Реальных размеров в корпусе: $248 + 362 = 610$
- У Шенгели (потенциальных размеров): 605 (!)
- Удивительно точная оценка при неверных посылках (ошибка в оценке допустимой стопности)!
- Заполненность ячеек в матрице потенциальности: 18,15 %

Зачем это нужно?

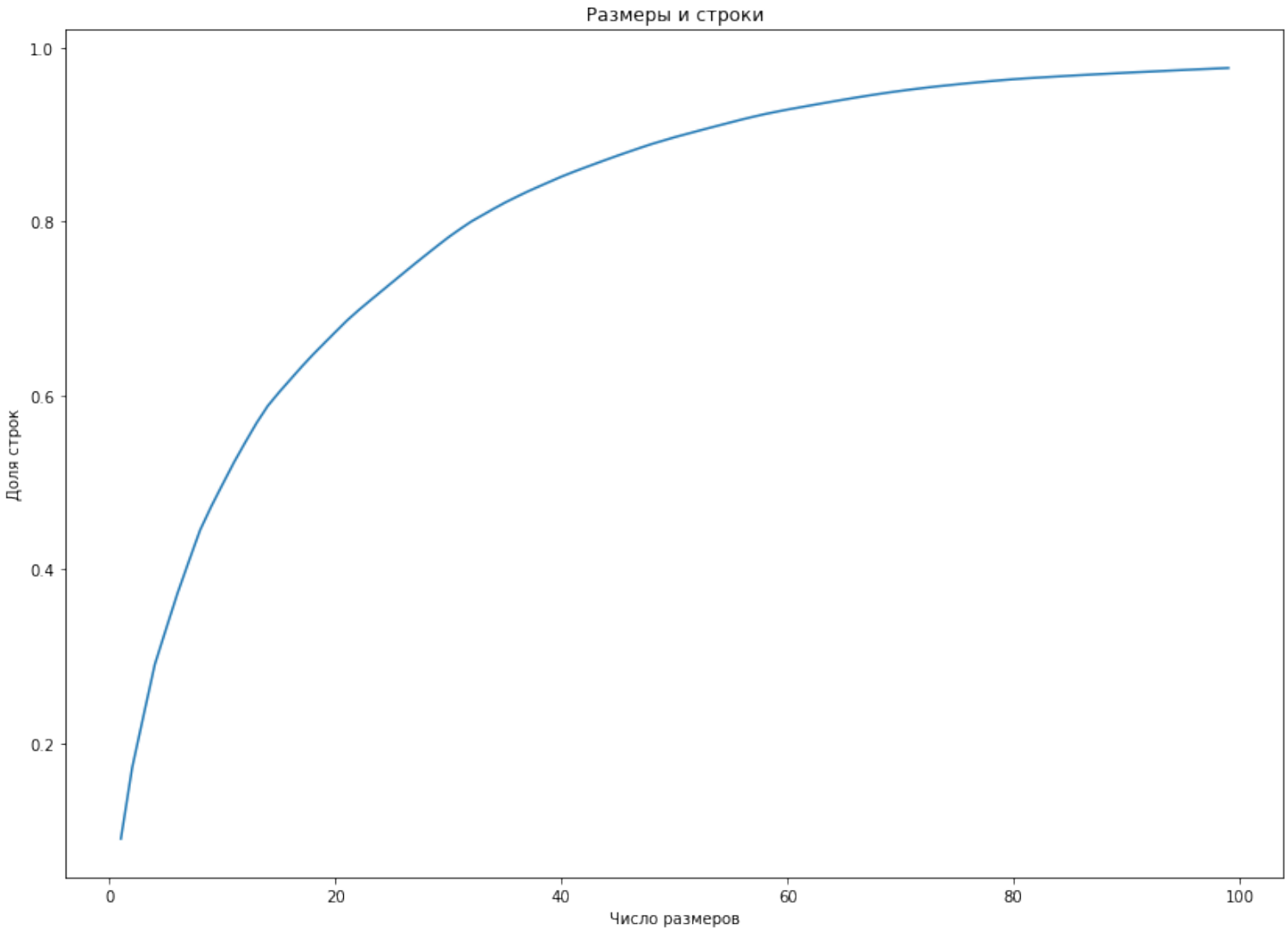
- Размеры — это информация.
- Ситуация обратная по сравнению с частотным словарем.
- В частотном словаре наиболее значимо то, что **частотно**.
- Наиболее информативны **редкие** поэтические размеры.
- Я4 не информативен.

Теория информации и стих

А что нам нравится и не нравится, — это определяется напластованием огромного множества впечатлений от всего прочитанного нами, начиная с первых детских стишков и до последних самых умных книг. Если новое стихотворение целиком похоже на то, что мы уже много раз читали, то оно ощущается как плохая, **скучная поэзия**; если оно решительно ничем не похоже на то, что мы читали, то оно ощущается как **вообще не поэзия**; хорошим нам кажется то, что лежит где-то посередине между этими крайностями, а где именно — определяет наш вкус, итог нашего читательского опыта. (М. Л. Гаспаров)

Скучная поэзия

50 размеров покрывают больше 95 % силлабо-тонических строк поэтического корпуса



Метрические раритеты в частотном окружении

Я5м В своих домах, и все тяжелым сном
Я1ж Я4ж Заснуло. И вот над этой темной бездной
Я5м От туч, их затмевавших, небеса,
Я6ж Уж полные звездами ночи, стали чисты...

В. А. Жуковский «Странствующий жид» (1851-1852)

Содержание

Распознавание размеров?

- Ключевое понятие силлабо-тоники — это ударение (ударный слог).
- Этот тезис выглядит как трюизм.
- Кажется, никто до сих пор никто не пытался его оспорить.
- Сама идея отрицания ключевой роли ударения выглядит абсурдно.

Стихи, порожденные символьной рекуррентной сетью

RNN, character based

И так же всё теперь слова.
Проснись ль скорей и воздух дал.
Но вот к конце весенних дней
И облачное на челе,
Спешит полутвердил я в небе,
И не сказал бы все врага,
И от меня в нем был лишь славой.

Обучающие данные

. . .
Заснул и я, но тяжек сон
Того, кто горем удручен.
Во сне я видел, что герой
Моей поэмы роковой
С полуобритой головой,
В одежде арестантских рот
Вдоль по Владимирке идет.
. . .

Специфика гуманитарных наук

- Обучающая выборка не содержит данных об ударении (!)
- Генеративная модель «угадывает» размер без ударения.
- Может ли классификатор угадывать размер без ударения?

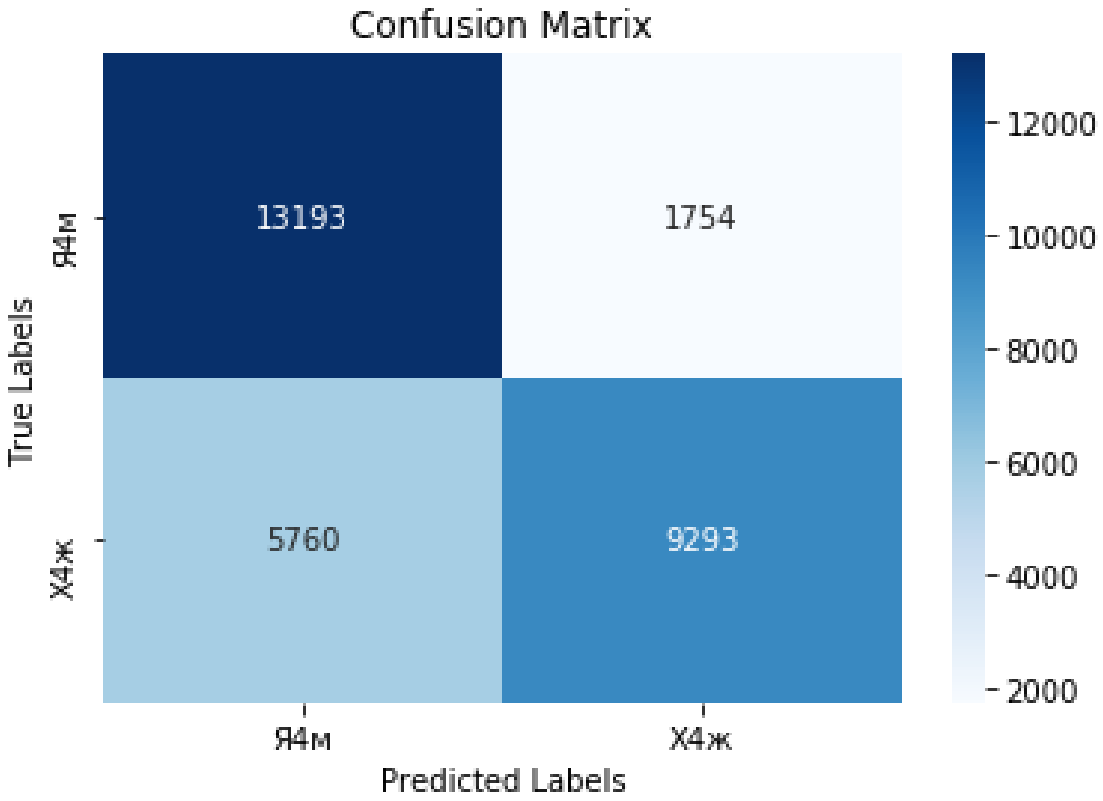
- Если **нет**, то наше представление о размере и ударении было верным.
- Если **да**, то ударение не обязательно для определения размера, что мы раньше назвали абсурдным.

Обучающие данные

Бинарная классификация

- 100 000 строк Я4м
- 100 000 строк Х4ж
- Я4м: «А сердцем больше щедр и благ», «Поверхность вод позолотит», «К ее пленительным устам»...
- Х4ж: ««Ничего... Жену спровадил», «Нам лишь чудо путь укажет», «О бананах долгоплодых»...

Оценка модели



0.7495

Почему так вышло?

- Самыми частотными сочетаниями букв в тексте наряду со служебными словами являются морфологические форманты (суффикс+окончание).
- Форманты различают части речи.
- Распределение частей речи в строке не случайны.
  Прилагательные предпочитают вторую половину строки, особенно третью стопу (. . .) Что касается наречий, то они предпочитают первую половину строки (Лингвистика стиха, с. 67)
- Сеть выучивает тенденции в распределении морфологических форм в строке, характерном для конкретного размера.
- Схожие наблюдения уже были сделаны на материале стилеметрического анализа (Sapkota, 2015)
  Models based only on affix and punct n-grams performed as well as models with all n-grams regardless of whether it was a single-domain or cross-domain authorship attribution task.

Другие важные для меня проекты и тексты

https://nevmenandr.github.io/