Статистика подслов

В. И. Арнольд

17 июля 2009 года

Аннотация

Найдя русское слово из 19 букв, содержащее 508 подслов, я задумался о поведении количеств подслов разных длин в разных текстах.

Эмпирические исследования, резюмируемые в настоящей статье, показывают, что эти распределения длин (и общие количества подслов) имеют для достаточно длинных текстов определённые асимптотики, мало зависящие от самих текстов.

Ключевые слова. «Евгений Онегин», «Шестое чувство», «Не станем пить». Статистическая физика русского языка. Гауссово распределение. Универсальность.

§1. Подслова данного текста

Рассмотрим набор n букв (вроде «Мой дядя самых честных правил» — в первой строке Онегина Пушкина, $n = 25$ букв.

Подсловами мы будем считать существительные, образованные из k букв данного текста. Примеры: «чехарда» ($k = 7$), «сапсан» ($k = 6$).

Целью настоящей статьи является эмпирическое исследование статистики подобных подслов (мы учитываем только нарицательные имена существительные именительного падежа в единственном числе, состоящие из $k \geqslant 4$ букв).

Пример 1. Сколько всего таких подслов имеет данный текст?

Я насчитал (за пару часов) $N = 370$ подслов для приведенной выше строки Онегина из $n = 25$ букв. Было бы интересно понять, насколько сильно числа подслов N различаются для разных текстов фиксированной длины n (а также насколько быстро растёт их максимальное число $N(n)$ по всем текстам длины n).

Я насчитал (за пару часов) для одного слова длины $n = 19$ пятьсот восемь подслов ($N = 508$), а для другого слова той же длины ($n = 19$) N оказалось равным 128 ($N = 128$).

Пример 2. Среди N подслов (данного текста) имеются подслова разных длин ($m = 4, m = 5,\ \ldots$):

$$N = N_4 + N_5 + \ldots$$

Вопрос заключается в том, как распределены длины подслов m, как ведут себя отношения

$$p_m = N_m/N$$

при росте длины m рассматриваемых подслов?

Было бы интересно также понять, сильно ли получающееся распределение $\{p_m\}$ зависит от исходного текста — если эти распределения для разных текстов получаются сходными, то возникает «универсальное» распределение, которое интересно было бы изучить.

Например, первые $n = 25$ букв Евгения Онегина можно было бы заменить первыми $n = 100$ буквами, или всею первой строфой, или даже всею первой главой.

Вопрос о том, доставят ли разные стихотворения разные распределения p_m также интересен — причём интересен и вопрос о различии подобных статистик для поэтических и для прозаических текстов.

Пример 3. Какова средняя длина подслова данного (n-буквенного) текста?

Средняя длина определяется здесь как среднее арифметическое длин всех подслов:

$$\hat{m} = \sum_m (mp_m) = \frac{1}{N} \sum_m (mN_m) \,.$$

Интересно было бы посмотреть, сильно ли различаются средние длины подслов разных исходных текстов? Интересно изучить и максимальное значение

$$\hat{M}(n) = \max \hat{m}$$

(где максимум берётся по всем отдельным строкам Онегина).

Поведение средних значений $\hat{M}(n)$ с ростом длины n исходного текста — также интересный объект исследования: будет ли $\hat{M}(n)$ неограниченно расти с ростом n, или же по мере роста n установится предельное значение $\hat{M}(\infty)$?

§2. Числа подслов

В этом разделе приведены эмпирические результаты моих подсчётов чисел подслов для следующих трёх исходных текстов (Пушкина, Гумилёва и Ахматовой) из n букв:

I. Мой дядя самых честных правил ($n \leqslant 25$);

II. Прекрасно в нас влюблённое вино ($n \leqslant 27$);

III. Не станем пить из одного стакана ($n \leqslant 27$).

В каждом из этих трёх случаев я перечислил за пару часов по нескольку сот подслов. Глядя на их списки, я легко вычислил для каждого слова те значения длины n исходного текста, для которых начальный отрезок длины n рассматриваемого текста содержит исследуемое подслово.

Например, в случае строки I (из Онегина) я получил следующие значения n:

$$n(\text{«а}\underline{\text{в}}\text{тор»}) \geqslant 23, \quad n(\text{«рад}\underline{\text{и}}\text{о»}) \geqslant 24, \quad n(\text{«со}\underline{\text{х}}\text{а»}) \geqslant 12, \quad n(\text{«чемода}\underline{\text{н}}\text{»}) \geqslant 17.$$

Здесь подчёркнуты критические буквы. Например, «соха» нуждается в букве «х», отсутствующей среди первых 11 букв изучаемой строчки (в которой буква «х» стоит на двенадцатом месте).

Число тех подслов W, для которых $n(W) \leqslant m$, и есть эмпирически найденное значение $N(m)$ (вычисленное для первых n букв изучаемой строки).

Выбирая $n = 10, 15, 20$ и 25, я получил в примерах I. «Онегин» Пушкина, II. «Шестое чувство» Гумилёва и III. «Из одного стакана» Ахматовой следующие числа $N(n)$:

Текст \ n	10	15	20	25
I	6	17	54	370
II	56	110	174	303
III	28	98	165	317

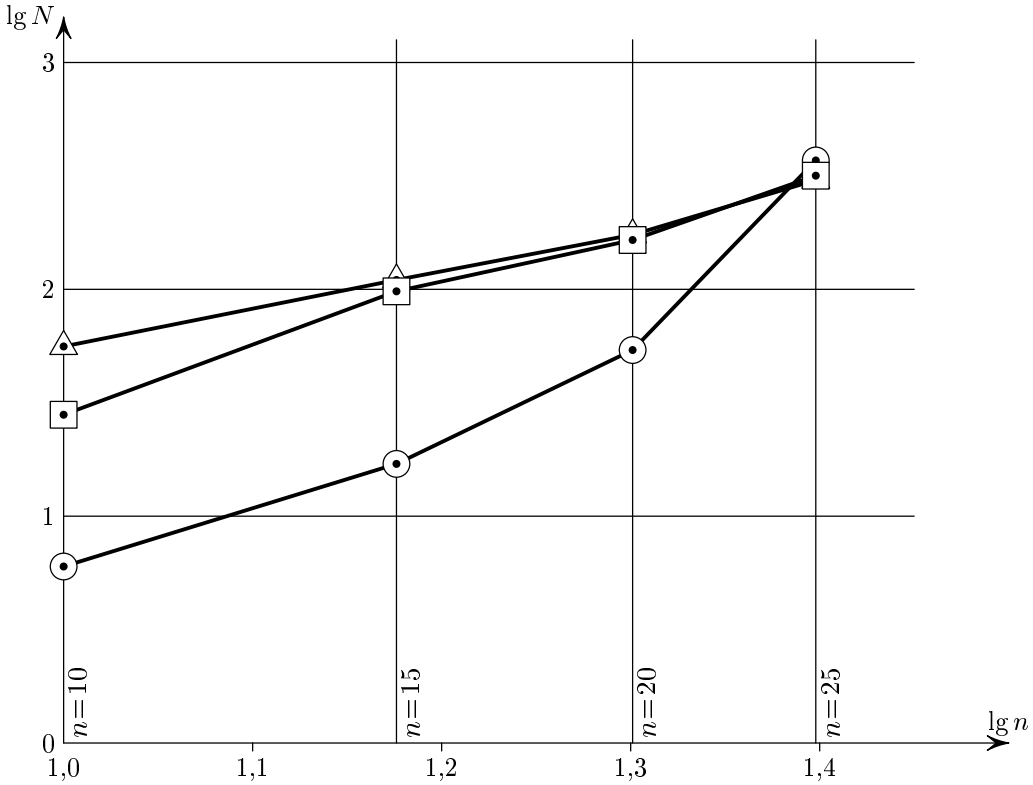
Чтобы понять поведение значения $N(n)$ с ростом n, я нарисовал графики трёх описанных функций на билогарифмической бумаге. Выбранные 4 значения n приводят к абсциссам

$$\lg 10 = 1{,}00; \quad \lg 15 \approx 1{,}18; \quad \lg 20 \approx 1{,}30; \quad \lg 25 \approx 1{,}40.$$

Поэтому указанные 3 графика проходят через точки с такими координатами на билогарифмической бумаге:

Текст \ $\lg n$	1,00	1,18	1,30	1,40
I	0,78	1,22	1,73	2,57
II	1,75	2,04	2,24	2,48
III	1,45	1,99	2,22	2,50

Каждый из трёх графиков напоминает прямую линию:



Аппроксимируя графики прямыми наклона p (которые соответствуют степенным

законам $N \sim Cn^p$), мы находим, соответственно:

$$\bigcirc \text{ I}: \quad p \approx \frac{257-78}{140-100} \approx 4{,}5;$$
$$\triangle \text{ II}: \quad p \approx \frac{248-175}{140-100} \approx 1{,}8;$$
$$\square \text{ III}: \quad p \approx \frac{250-145}{140-100} \approx 2{,}6.$$

Замечание. Эмпирические степенные зависимости ($N \sim \mathrm{Const}\, n^p$) кажутся разумными приближениями, но сотни дальнейших экспериментов (со значительно большим, чем 25, значениями длин текстов n) доставляют меньшие наклоны p, вплоть до $p \sim 1$ (что соответствовало бы уже линейному приближению $N \approx \mathrm{Const}\, n$).

Чтобы выбрать, следовало бы систематически изучить более длинные тексты ($N \approx 1000$?). Но перечисление всех $N(n)$ подслов таких текстов становится при больших n очень трудоёмким. Поэтому я использовал с этой целью другую технологию, перечисляя гораздо меньшие количества $N_m (\ll N(n))$ подслов фиксированной длины ($m = 4, 5, \ldots$).

Действительно, список всех 4-буквенных подслов данного текста составить не так уж трудно. Обозначим их число через N'_4. Предположим, что эмпирическое перечисление этих подслов (за пару часов работы) доставило только $N_4 < N'_4$ подслов длины 4. Можно предположить, что вероятность включить в эмпирическое перечисление примерно одинакова для подслов из 4 букв и для всех подслов:

$$\frac{N_4}{N'_4} \approx \frac{N}{N'}$$

Тогда мы получим из неполного числа N эмпирически перечисленных подслов ожидаемое бóльшее число всех подслов: $N' \sim N(N'_4/N_4)$. Время вычисления этого ожидаемого числа N' значительно меньше, чем время исчерпывающего перечисления всех подслов (для текста длины $n \sim 1000$).

Получающиеся этим методом значения чисел подслов указаны в следующем разделе.

Думаю, что для 25 букв первой строки Онегина (где я нашёл $N = 370$ подслов), общее число подслов $N \approx 450$.

§3. Распределение длин подслов

Перечисляя подслова длин $m = 4, 5$ и 6 (в составленных списках N подслов любых длин в данном тексте), я посчитал в своих списках, составленных по текстам I, II и III длины n, следующие количества $N(m)$ подслов из m букв:

Текст \ m	4	5	6	N
I	152	136	46	370
II	100	102	58	303
III	81	105	72	317

Переходя от долей $p_m = N_m/N$ к процентным отношениям $100p_m$, мы получаем из предыдущей таблицы следующие процентные отношения (чисел подслов длин 4, 5 и 6):

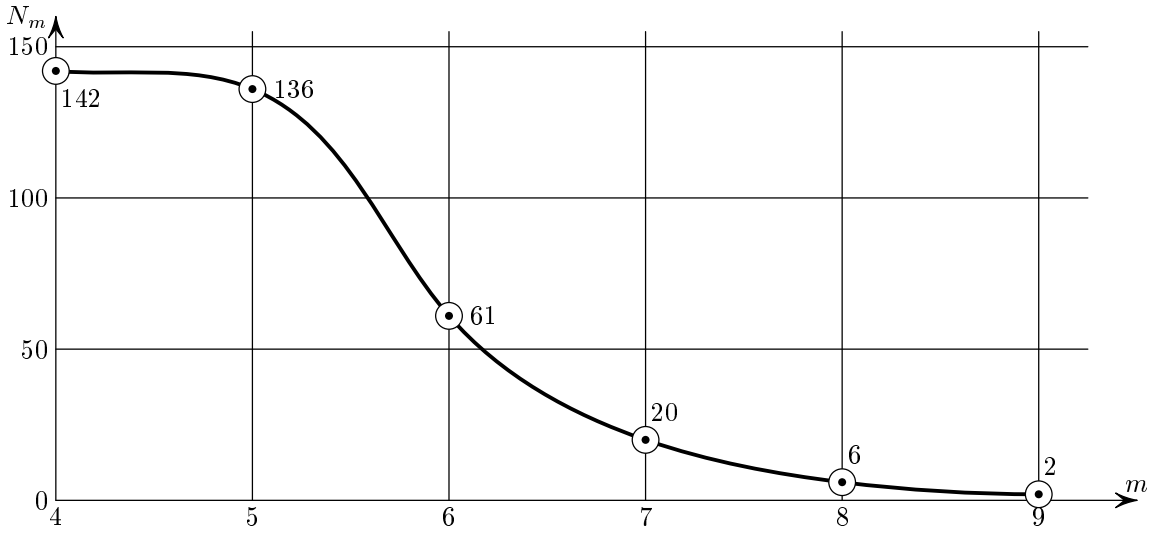
Текст \ m	4	5	6
I	40,1	30,7	12,4
II	32,7	33,4	19,0
III	25,5	33,1	22,7

Соответствующие длины подслов, $\hat{m} = \sum_m (mp_m)$, составляют

Текст	I	II	III
$\hat{m}$	4,23	4,11	4,92

Эти таблицы показывают, что бóльшая часть подслов имеет длины 4 или 5, так что средняя длина подслова всегда лежит недалеко от числа 4,5.

Замечание. Наряду со статистикой длин $m = 4, 5$ и 6 распределения $\{p_m\}$ всюду напоминают сдвинутое Гауссово распределение. Например, для случая I (начало «Онегина») распределение длин такое:



Предполагая квадратичное (по Гауссу) поведение логарифмов чисел слов

$$\lg N_m \approx am^2 + bm + c,$$

мы получаем

$$16a + 4b + c = \lg 152 \approx 2{,}18$$
$$25a + 5b + c = \lg 136 \approx 2{,}12$$
$$36a + 6b + c = \lg 46 \approx 1{,}66$$

Эти три точки параболы доставляют значение

$$9a + b \approx 0{,}06; \quad 11a + b \approx -0{,}46;$$

откуда получается

$$2a \approx -0{,}52, \quad a \approx -0{,}26, \quad b \approx 2{,}40.$$

Высшая точка параболы соответствует длине слова

$$m_* = -\frac{b}{2a} \approx \frac{2{,}40}{0{,}52} \approx 4{,}6$$

Таким образом, все эти вычисления приводят к гипотезе, что при $n \to \infty$ средние длины подслов текста длины n стремятся к универсальному (не зависящему от исходного текста) пределу $M \approx 4{,}5$.

§4. Влияние длины исходного текста на распределение длин его подслов

Заменяя в предыдущих вычислениях полные списки подслов исходного текста его частями, состоящими из подслов начальных n-буквенных отрезков текста, я получил для чисел подслов длин $m = 4, 5, 6$ отрезков текста длин $n = 15, 20, 25$ следующие значения:

	m \ n	15	20	25
I	4	14	26	141
	5	3	19	143
	6	0	3	41
	N	17	54	370
II	4	52	68	100
	5	40	57	102
	6	13	30	57
	N	110	174	303
III	4	48	59	81
	5	37	54	105
	6	14	30	72
	N	98	165	317

Заменяя числа N_m подслов длины m их процентной долей $100p_m$ (где $p_m = N_m/N$), мы получаем следующую таблицу процентных долей (подслов длин m в текстах длин n):

	m \ n	15	20	25
	4	82	48	38
I	5	17	35	39
	6	0	5	12
	4	47,3	39,1	33,0
II	5	36,4	32,8	33,7
	6	11,8	17,2	18,8
	4	49,0	35,8	25,6
III	5	37,7	32,7	33,1
	6	14,3	18,2	22,7

Эти распределения подслов длин m (среди подслов текстов длин n) кажутся сходящимися при $n \to \infty$ к некоторому универсальному распределению, для которого ожидаемые (приближённые) пропорции чисел слов длины $m = 4, 5$ и 6 составляют

$$p_4 \approx 0{,}32; \quad p_5 \approx 0{,}35; \quad p_6 \approx 0{,}18.$$

При этом заметно, что значения p_4, по-видимому, убывают с ростом длины n текста, а остальные значения ($p_{>4}$), по видимому, растут вместе с n.

Предполагая параболу Гаусса

$$\lg p_m \approx am^2 + bm + c,$$

мы находим на ней 3 точки

$$16a + 4b + c \approx -0{,}50,$$
$$25a + 5b + c \approx -0{,}45,$$
$$36a + 6b + c \approx -0{,}75.$$

Из этих значений мы находим

$$9a + b \approx 0{,}05, \quad 11a + b \approx -0{,}30, \quad 2a \approx -0{,}35.$$

Таким образом, наши вычисления дали бы, в предположении Гауссовости,

$$a \approx -0{,}18, \quad 9a \approx -1{,}58, \quad b \approx 1{,}63.$$

Для таких значений параметров распределения Гаусса среднее значение длины подслова составило бы

$$m_* = -\frac{b}{2a} \approx \frac{1{,}63}{0{,}35} \approx 4{,}7.$$

Полученная средняя длина подслова недалека от оценок предыдущих разделов, эмпирически подтверждая и гипотетическое значение средней длины подслова, $M \sim 4{,}5$, и предполагавшуюся выше Гауссовость распределения длин подслов.