УДК 517.97 : 532.526

О ПРЕДЕЛЬНОМ РАЗМЕРЕ СЛОВАРЯ И ФРАКТАЛЬНОЙ РАЗМЕРНОСТИ МЕТАКНИГИ М.Е. САЛТЫКОВА-ЩЕДРИНА

А.А. Кретов kretov@rgph.vsu.ru[1]
М.В. Половинкина polovinkina-marina@yandex.ru[2]
И.П. Половинкинpolovinkin@yandex.ru. [1,3]
Н.А. Касымова kasymova.natasha@mail.ru[1]

[1]*Воронежский государственный университет,*
[2]*Воронежский государственный университет инженерных технологий,*
[3]*Белгородский государственный национальный исследовательский университет*

Аннотация. *В работе описываются возможности измерения таких характеристик авторских корпусов текстов, как предельный размер словаря писателя и фрактальная размерность его метакниги. Рассматривается проблема практического расчета фрактальной размерности. Приводятся результаты расчетов для метакниги М.Е. Салтыкова-Щедрина.*

Ключевые слова: *закон Хипса, самоподобие, фрактальность языка, фрактальная размерность.*

Введение

Во многих сферах научных исследований может быть применен аппарат нелинейной динамики. В частности, это можно сказать о принципе самоподобия и понятии фрактала. Можно сказать, что понятие фрактала в математике и физике закрепилось устойчиво. В других областях обнаружение эффектов самоподобия и возможность использования инструментов фрактальной теории достаточно разрозненны, хотя база выявленных фактов достаточно обширна. Мы предлагаем рассмотреть некоторые из достижений современной лингвистики с точки зрения теории фракталов. Фрактальные (рекурсивно-самоподобные) проявления в языке были замечены в лингвистических исследованиях (см., например, [1-4]). В основном речь идет о фиксации и словесном описании самоподобия в языке. Однако есть все основания рассматривать количественные характеристики фрактальности текстов.

1. Фрактальная размерность текста метакниги и способ ее оценки.

В работе [5] предлагается уточнение закона Хипса (со ссылкой на [6]), согласно которому количество различных, уникальных слов, лемм (N), как функция от общего количества слов (словоупотреблений) в метакниге (M), имеет степенной порядок роста $\Theta(M^{\alpha})$, где $\alpha \in (0,1)$. Далее предлагается рассматривать закон Хипса не как асимптотическую оценку, а как точную формулу с переменным показателем α и переписать его в виде

$$\alpha = \alpha(M) = \ln N / \ln M .$$

Это является основанием обратиться к аппарату, развитому в теории фракталов. В книге [7] описан следующий поход к понятию фрактальной размерности. Введем в пространстве R^d совокупность конгруэнтных «атомарных» множеств, имеющих топологическую размерность d. Это множество либо d-мерных шаров, либо d-мерных кубов. Для определенности будем считать, что это шары. Пусть фрактальный объект находится в пространстве R^d. Зафиксируем достаточно малый радиус $l > 0$. Покроем целиком фрактальный объект шарами радиуса l. Предположим, что для этого потребовалось как

минимум $N = N(l)$ шаров. Число

$$\alpha_0 = -\lim_{l \to 0} (\ln N / \ln l) = \lim_{l \to 0} (\ln N / \ln(1/l))$$

называется фрактальной размерностью рассматриваемого объекта. В форме (2) это определение едва ли подойдет для характеристики текста, поскольку мы не можем устремлять к нулю размер атомарного множества, которым естественно считать слово (словоупотребление). Придется его немного изменить с целью приспособить к нашим нуждам.

В обозначениях [5] положим

$$l = 1/M .$$

Можно интерпретировать равенство (3) следующим образом. Считая словоупотребление «атомарным кирпичиком» для рассматриваемого текста, мы определяем его размер, соизмеряя этот «кирпичик» с самим же текстом, так как, собственно, его больше нечем измерить. Иными словами, за размер «атома» мы принимаем долю, занимаемую им в целом. Под мощностью же покрытия текста мы понимаем количество уникальных слов (лемм), словоупотребления которых составили весь текст. Далее по определению положим

$$\alpha_0 = -\lim_{l \to 0} (\ln N / \ln l) = \lim_{M \to +\infty} (\ln N / \ln M) = \lim_{M \to +\infty} \alpha(M),$$

а число α_0, определенное формулой (4), назовем фрактальной размерностью текста.

Практическое вычисление числа α_0 по формуле (4), конечно, невозможно. В формуле (4) предполагается, что объем текста M, понимаемый как количество словоупотреблений в нем, может принимать сколь угодно большие значения. Если речь идет о тексте некоторого произведения, то, разумеется, это не так. Авторы работы [5] вводят понятие метакниги писателя как объединения всех текстов, написанных этим писателем. Если писатель достаточно плодовит, то такая концепция позволяет считать, что $M \to +\infty$, хотя при практическом вычислении все равно приходится ограничиваться имеющейся длиной метакниги для вычисления приближенного значения α_0.

Нижняя оценка фрактальной размерности метакниги может быть получена из следующих соображений. На основе эмпирических данных произведем аппроксимацию функции, выражающей зависимость величины словаря от величины метакниги. Пользуясь полученной зависимостью, с помощью экстраполяции определим такую величину метакниги, при превышении которой приращение величины словаря будет пренебрежимо мало. Найдем соответствующий предельный объем словаря и вычислим величину (1) для найденных значений.

Немного видоизмененный подход может состоять в следующем. Обратимся к важной характеристике мета-книги, называемой "коэффициентом лексического разнообразия" (КЛР, англ. lexical diversity, LD) – количественная характеристика текста, отражающая степень богатства словаря при построении текста заданной длины. В самом простом варианте LD вычисляется как отношение числа отдельных лексических единиц словаря (лемм, англ. types) к количеству их употреблений в тексте (словоформ, «текстовых слов», англ. tokens) (type/token ratio) [https://ru.wikipedia.org/wiki/Коэффициент лексического разнообразия]. Для такого способа вычисления принято обозначение TTR. TTR предположительно был введен в научный обиход в 1957 году в работе специалиста по лингводидактике М. Темплина (см., напр., [8]). Вычисление LD в виде TTR подвергается критике за то, что при этом "не учитывается влияние длины текста", поскольку при увеличении длины текста величина словаря растет медленнее, а значит TTR будет уменьшаться и стремиться к нулю. Однако для наших целей именно это качество TTR полезно. Можно считать предельным размером словаря такое значение этого размера, при котором КЛР становится пренебрежимо малым. В связи с этим требуется уточнить, что понимается под "малостью" как приращения словаря, так и КЛР. Здесь возникает и

проблема увязать это понятие малости с выбором модели тренда и как следствие – способа экстраполяции тренда.

2. Верхняя оценка фрактальной размерности метакниги М.Е. Салтыкова-Щедрина

В качестве примера применения изложенных выше соображений мы рассмотрели 20 произведений М.Е. Салтыкова-Щедрина разного объема, охватывающие более-менее равномерно отрезок времени в 20 лет. При этом сознательно брались тексты разного размера, чтобы иметь дело с наиболее сложным случаем прироста новых слов. Нам пришлось совершить 19 шагов, на каждом из которых метакнига наращивалась посредством конкатенации текста очередного произведения, вычислялся ее текущий размер, равный количеству словоупотреблений, а также осуществлялись лемматизация, соответствующее наращивание словаря и вычисление его текущего размера. Лемматизация осуществлялась с помощью размещенного в свободном доступе морфологического анализатора русского языка MyStem, разработанного Ильей Сегаловичем в компании "Яндекс". На основе расчетов, произведенных с этим корпусом текстов (метакнигой), мы пришли к верхней оценке фрактальной размерности метакниги М.Е. Салтыкова-Щедрина, равной 0,74307.

3. Предельный размер словаря и нижняя оценка фрактальной размерности метакниги М.Е. Салтыкова-Щедрина

Для верхней оценки нам понадобились лишь конечные значения размера метакниги и размера словаря. Для нижней оценки понадобилась фиксация всех промежуточных пар значений после каждой конкатенации. Эти данные приведены в табл. 1.

Табл. 1. Прирост новых слов и покрываемого ими текста.
Tabl.The growth of new words and the text covered by them

№	Год	Текст	Длина Δ_M	Слов Δ_N	КоЛеР Δ_N / Δ_M	ДлКум M	СлКум N	КоЛеР-кум $Y_{TTR} = N/M$
1	1857	Губернские очерки	144531	13124	0,09081	144531	13124	0,091
2	1862	Сатиры в прозе	47017	7569	0,16098	191548	15746	0,082
3	1863	Невинные рассказы	30252	5439	0,17979	221800	16875	0,076
4	1870	История одного города	58424	10172	0,17411	280224	21013	0,075
5	1872	Дневник провинциала в Петербурге	95263	10329	0,10843	375487	23879	0,064
6	1873	В больнице для умалишенных	20814	4245	0,20395	396301	24482	0,062
7	1873	Господа ташкентцы	96683	10960	0,11336	492984	26753	0,054
8	1873	Современная идиллия	96524	11234	0,11639	589508	28996	0,049
9	1874	Помпадуры и помпадурши	83689	9718	0,11612	673197	30217	0,045
10	1874	Тихое пристанище	25850	5012	0,19389	699047	30632	0,044

11	1876	Благонамеренные речи	165706	14520	0,08763	864753	33268	0,038
12	1878	В среде умеренности и аккуратности Господа Молчалины	205115	14740	0,07186	1069868	35553	0,033
13	1879	Убежище Монрепо	50227	7406	0,14745	1120095	36141	0,032
14	1880	Господа Головлевы	90792	9947	0,10956	1210887	37179	0,031
15	1881	ЗаРубежом	81823	10066	0,12302	1292710	38248	0,030
16	1882	Письма к тетеньке	76159	9271	0,12173	1368869	39012	0,028
17	1886	Недоконченные речи	48004	7264	0,15132	1416873	39531	0,028
18	1886	Сказки	81254	10091	0,12419	1498127	40467	0,027
19	1887	Мелочи жизни	107365	11385	0,10604	1605492	41450	0,026
20	1889	Пошехонская старина. Начало	74711	9761	0,13065	1680203	42256	0,025

В табл. 1 N – текущее значение размера словаря; $^{\Delta}N$ – приращение словаря, то есть количество новых уникальных слов при присоединении очередного текста к метакниге; M – текущее значение размера метакниги; $^{\Delta}M$ – приращение размера метакниги, то есть количество словоупотреблений в присоединяемом к метакниге тексте; Y_{TTR} – текущее значение TTR (КЛР).

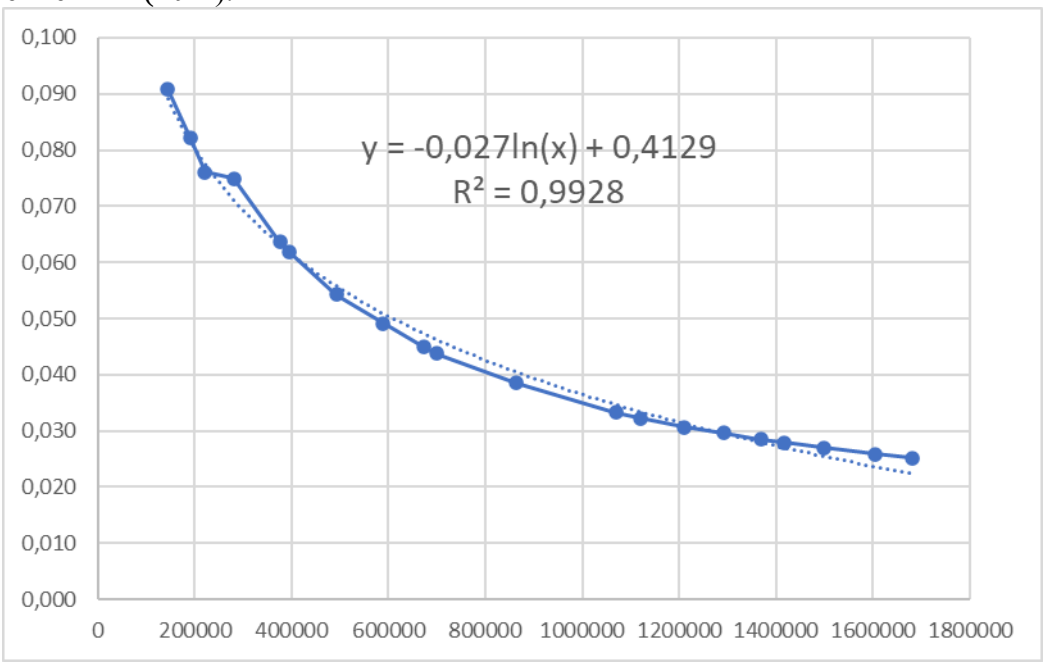

Рис.1. Динамика КЛР в корпусе художественной прозы М.Е. Салтыкова-Щедрина при присоединении к корпусу новых произведений.

Fig 1. LD dynamics in the corpus of fiction by M.E. Saltykov-Shchedrin when joining the corpus new works.

Возникает естественный вопрос об адекватном моделировании тренда изменения КЛР с увеличением корпуса произведений писателя. Выбор средств моделирования, как известно, зависит от целей моделирования. В нашем случае этой целью является определение предельного размера словаря. Здесь возникает еще одна задача: указать формализованные признаки достижения предельного размера словаря. В качестве таковых можно предложить близость к нулю приращения словаря при включении в корпус текста очередного произведения или близость к нулю КЛР. Совершенно ясно, что величина КРЛ должна стремиться к нулю при неограниченном увеличении корпуса, но принимать нулевое значение не может, поскольку величина размера словаря всегда положительна. В связи с этим требуется уточнить, что понимается под «малостью» как приращения словаря, так и КЛР. Здесь возникает и проблема увязать это понятие малости с выбором модели тренда и как следствие способа экстраполяции тренда.

Имеются многочисленные попытки построения эмпирических формул для выражения зависимости объема словаря от объема текста, как и зависимости КРЛ от объема текста. Наиболее подходящей в агрегированном смысле считается аппроксимация по степенному закону Ципфа, известному также как закон «аллометрического» или «постоянного относительного роста»:

$$КЛР = C\,x^{\gamma}$$

где $\gamma < 0$, x – накопленный размер текста корпуса. При таком моделировании тренда мы, конечно, не получим нулевого значения КЛР, что соответствует реальности. Поэтому мы можем считать, что рост словаря пренебрежимо мал, когда КЛР пренебрежимо мал. Что это означает, подчеркиваем, подлежит уточнению. Кроме проблемы уточнения «малости» есть еще одна проблема. Согласно Ю.А. Тулдаве [9, с. 99] при больших размерах текста прогнозирование тренда КЛР с помощью закона Ципфа дает значительные погрешности (завышенные оценки).

Мы предлагаем несколько иной путь. Выберем в качестве линии тренда логарифмическую зависимость (см. рис.). Более точно, мы выбираем логарифмические и постоянные функции в качестве базисных, а функцию зависимости КЛР от объема текста ищем в виде линейной комбинации базисных функций. Коэффициент правдоподобия в таком случае тоже очень высок. Зато такая функция имеет нуль. Значение размера текста при этом мы можем считать соответствующим предельному размеру словаря. Приравняем нулю функцию тренда и решим уравнение

$$0{,}4129 - 0{,}027 \ln M = 0.$$

Пусть M_0 – корень этого уравнения. Легко видеть, что

$$\ln M_0 = 0{,}4129 / 0{,}027 \approx 15{,}29259 \quad M_0 \approx e^{15{,}29259} \approx 4380146.$$

Итак, исходя из выбранного способа моделирования, мы заключаем, что размер текста корпуса, при котором достигается предельный размер словаря М.Е. Салтыкова-Щедрина, составляет 4 380146 слов. Ясно, что это некоторая приближенная оценка.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

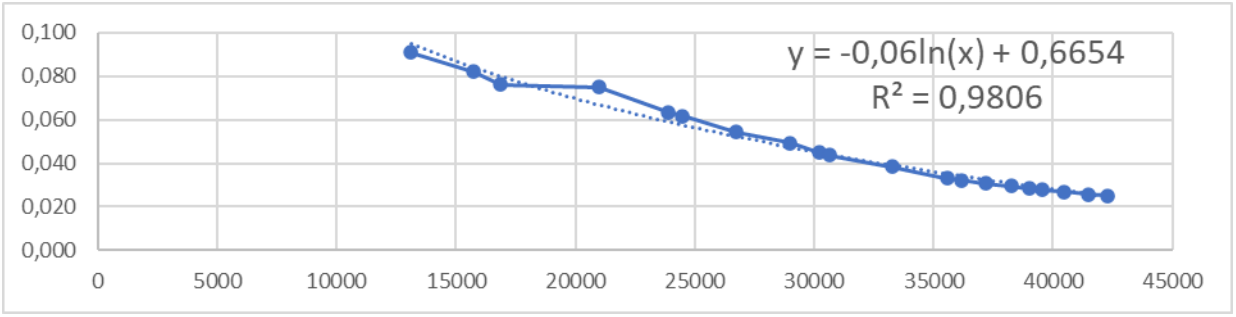

Рис. 2. Зависимость КЛР от размера словаря М.Е. Салтыкова-Щедрина

Fig. 2. Dependence of TTR on the size of M.E. Saltykov-Shchedrin's dictionary

Теперь мы должны найти предельный размер словаря. Пойдем тем же путем. В качестве линии тренда выберем логарифмическую зависимость (см. рис.) и приравняем нулю функцию тренда. Пусть N_0 – корень уравнения

$$0,6654 - 0,06 \ln N = 0.$$

Тогда, очевидно,

$$\ln N_0 = 0,6654 / 0,06 \approx 11,09, \quad N_0 \approx e^{11,09} \approx 65512.$$

Итак, оценка предельного размера словаря М.Е. Салтыкова-Щедрина (с необходимым замечанием об учете выбранного метода моделирования) составляет «прогнозно» 65512 слов.

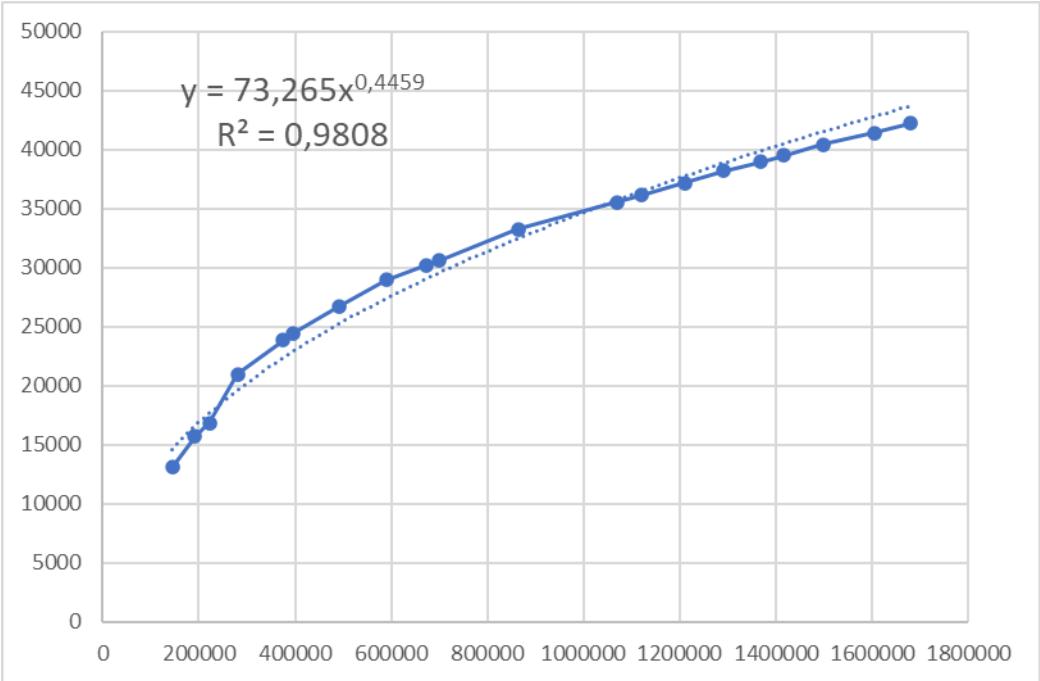

Рис. 3. Зависимость размера словаря от размера корпуса текстов М.Е. Салтыкова-Щедрина

Fig. 3.The dependence of the size of the dictionary on the size of M.E. Saltykov-Shchedrin's texts

Есть еще одна проблема – проблема проверки достоверности полученных прогнозов. Классический способ сравнения приближенного решения с точным решением или с экспериментальными данными применен быть не может по причине отсутствия таковых. Здесь нам доступны лишь косвенные способы проверки. Попробуем вновь воспользоваться вариантом закона Ципфа, но теперь для описания зависимости размера словаря от размера текста:

$$N = A M^{\beta},$$

где N – размер словаря, M – размер текста, $0 < \beta < 1$. По данным табл. устанавливается степенная зависимость вида (см. рис. 4):

$$N = 73,265 M^{0,4459}.$$

Подставив в эту формулу значение $M_0 \approx e^{15,29259} \approx 4380146$, мы получим значение 67 040 слов как оценку для предельного размера словаря. Это значение отличается от полученного ранее как нуля логарифмической функции тренда КЛР. Однако относительная погрешность составляет

$$\frac{67040 - 65512}{65512} \times 100\% \approx 2{,}33\%,$$

что, на наш взгляд, вполне приемлемо. Осталось только принять окончательное решение о прогнозе предельного размера словаря и размера соответствующего размера корпуса.

Произведя традиционные округления, приходим к следующим прогнозам:

- предельный размер словаря М.Е. Салтыкова-Щедрина находится в промежутке 65.512 – 67.040 слов,

- размер текста, при котором достигается предельный размер словаря М.Е. Салтыкова-Щедрина, составляет 4.380.146 слов.

Теперь мы можем вычислить нижнюю оценку фрактальной размерности метакниги М.Е. Салтыкова-Щедрина:

$$\alpha_0 \approx \frac{\ln 65512}{\ln 4380146} \approx 0{,}725188 .$$

Таким образом, фрактальная размерность метакниги М.Е. Салтыкова-Щедрина, составленной из его 20 произведений, может быть заключена в промежуток [0,725188; 0,74308].

Отметим, что ранее примененный здесь метод использован в [10].

Список литературы

1. Кретов А. А. Основы лексико-семантической прогностики. Монография / А. А. Кретов. – Воронеж: Изд-во ВГУ, 2006. – 404 с. [«Библиотека лингвистической прогностики». Том 1.]
2. Кретов А. А. Русское слово как самоподобная рекурсивная структура / А. А. Кретов, И. Е. Воронина // Лингвистика на исходе XX века: итоги и перспективы: сб. науч. труд. – М.: Филология, 1995. – Т. I. – С. 269 271.
3. Кретов А.А. Фрактальность в русском языке / А. А. Кретов // Русское национальное сознание в его языковом воплощении: прошлое, настоящее, будущее. XXX Распоповские чтения : материалы Международной конференции, Воронеж, 2-4 марта 2012 г. / [под ред. Л.М. Кольцовой] ; Воронежский государственный университет. – Воронеж : Издательско-полиграфический центр Воронежского государственного университета, 2012, С.138-147.
4. Петряков Л. Д. Методологические перспективы фрактальной семантики / Л. Д. Петряков // Известия вузов. Серия «Гуманитарные науки». – 2017. – 8 (2) – С. 148–153.
5. Bernhardsson S. The meta book and size-dependent properties of written language / S. Bernhardsson, L. E. Correa da Rocha, P. Minnhagen // New Journal of Physics. – 2009. – 11. – 123015 (15pp). Online at http://www.njp.org/ doi:10.1088/1367-2630/11/12/123015
6. Heaps H. S. Information Retrieval: Computational and Theoretical Aspects / H. S. Heaps – New York: Academic Press, 1978.
7. Mandelbrot B. B. The Fractal Geometry of Nature / B. B. Mandelbrot. – San Francisco: W.H. Freeman, 1982. – 468 p.
8. Torruella J. and Capsada R Lexical Statistics and Tipological Structures: A Measure of Lexical Richness / J. Torruella, R. Capsada // Procedia - Social and Behavioral Sciences. – 2013. – 95. – pp. 447–54.
9. Тулдава Ю.А. Проблемы и методы квантитативно-системного исследования лек-сики / Ю.А. Тулдава. -— Таллин: Валгус, 1987. — 204 с.
10. A A Kretov, M V Polovinkina, I P Polovinkin and M V Lometc On some concepts of nonlinear dynamics suitable for use in linguistics 2021 *J. Phys.: Conf. Ser.* 1902 doi:10.1088/1742-6596/1902/1/012075