О предельном размере словаря и фрактальной размерности метакниги А.Ф. Писемского

А.А. Кретов kretov@rgph.vsu.ru[1]
М.В. Половинкина polovinkina-marina@yandex.ru[2]
И.П. Половинкин polovinkin@yandex.ru[1,3]
А.В. Трофимова alina.trofimova01@mail.ru[1]

[1] Воронежский государственный университет
[2] Воронежский государственный университет инженерных технологий
[3] Белгородский государственный национальный исследовательский университет (БелГУ)

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 А.Ф. Писемского.*

Ключевые слова: *закон Хипса, самоподобие, фрактальность языка, фрактальная размерность, реальный и предельный размер словарного запаса А.Ф. Писемского.*

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 $\Theta\left(M^{\alpha}\right)$, 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.

$$\alpha = \alpha(M) = \ln N / \ln M . \qquad (1)$$

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

$$\alpha_0 = -\lim_{l \to 0} (\ln N / \ln l) = \lim_{l \to 0} (\ln N / \ln(1/l)) \qquad (2)$$

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

$$l = 1/M . \qquad (3)$$

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

$$\alpha_0 = -\lim_{l \to 0} (\ln N / \ln l) = \lim_{M \to +\infty} (\ln N / \ln M) = \lim_{M \to +\infty} \alpha(M), \quad (4)$$

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). 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 А.Ф. Писемского

В качестве примера применения изложенных выше соображений мы рассмотрели 20 произведений А.Ф. Писемского разного объема, охватывающие более-менее равномерно отрезок времени в 30 лет. При этом сознательно отбиралось 20 доступных текстов максимального размера, чтобы обеспечить наибольшую достоверность данных о приросте 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 А.Ф.Писемского, равной 0,726969.

3. Предельный размер словаря и нижняя оценка фрактальной размерности метакниги А.Ф. Писемского

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.

Таблица

Прирост новых слов и покрываемого ими текста.

Год 18__	Текст	Длина текста Δ M	Слов в тексте Δ N	Кумулят. длина корпуса текстов M	Кумулят. размер словаря автора N	КуКЛР Y_{TTR}

50	Тюфяк	54097	5263	54097	5263	0,0973
51	Сергей Петрович Хозаров и Мари Ступицына	40955	4462	95052	7000	0,0736
51	Комик	21366	3321	116418	7979	0,0685
53	Леший	13309	2524	129727	8836	0,0681
54	Фанфарон	17723	3003	147450	9505	0,0645
55	Виновата ли она?	23416	3163	170866	10081	0,0590
55	Плотничья артель	16637	3174	187503	11004	0,0587
58	Тысяча душ	136097	10673	323600	15279	0,0472
58	Боярщина	45694	4673	369294	15879	0,0430
59	Горькая судьбина	16052	2643	385346	16304	0,0423
61	Старческий грех	25958	4650	411304	17083	0,0415
63	Взбаламученное море	121661	10292	532965	19940	0,0374
65	Русские лгуны	51076	4049	584041	20419	0,0350
69	Люди сороковых годов	228215	12490	812256	23446	0,0289
71	В водовороте	147362	8056	959618	24455	0,0255
73	Подкопы	35756	2746	995374	24693	0,0248

73	Ваал	21946	3301	1017320	24971	0,0245
75	Просвещенное время	14091	2279	1031411	25133	0,0244
77	Мещане	95072	7922	1126483	26041	0,0231
80	Масоны	226047	13253	1352530	28652	0,0212

В табл. N – текущее значение размера словаря; ΔN – приращение словаря, то есть количество новых уникальных слов при присоединении очередного текста к метакниге; M – текущее значение размера метакниги; ΔM – 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 А.Ф. Писемского при присоединении к корпусу новых произведений.

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. 1). 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

$$0{,}3239 - 0{,}022 \ln M = 0$$

Пусть M_0 – корень этого уравнения. Легко видеть, что

$$\ln M_0 = 0{,}3239 / 0{,}022 \approx 14{,}72273,$$

$$M_0 \approx e^{14{,}72273} \approx 2477418$$

Итак, исходя из выбранного способа моделирования, мы заключаем, что размер текста корпуса, при котором достигается

предельный размер словаря А. Ф. Писемского, составляет 2 477 418 слов. Ясно, что это некоторая приближенная оценка.

4. Результаты и их обсуждение

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

$$0.4424 - 0.041 \ln N = 0$$

Тогда, очевидно,

$$\ln N_0 = 0{,}4424 / 0{,}041 \approx 10{,}79024,$$

$$N_0 \approx e^{10{,}79024} \approx 48545$$

Итак, оценка предельного размера словаря А.Ф. Писемского (с необходимым замечанием об учете выбранного метода моделирования) составляет «прогнозно» 48545 слов.

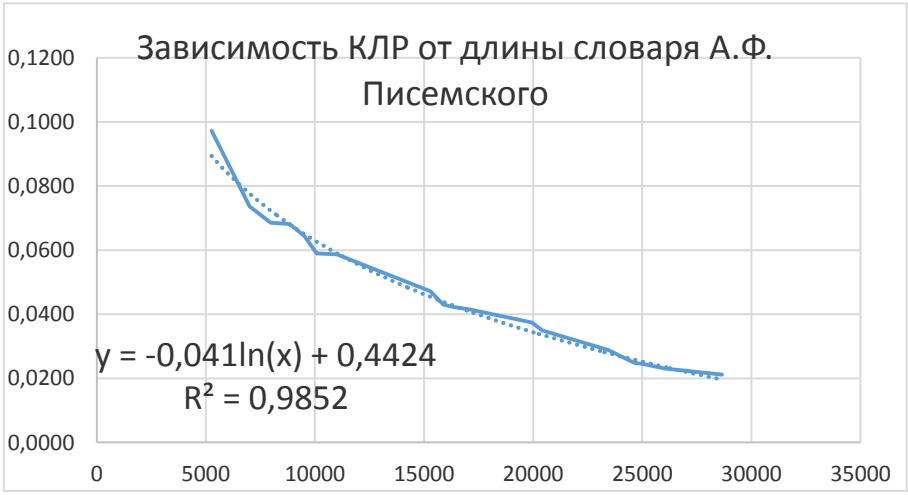

Рис. 2. Зависимость КЛР от размера словаря А.Ф. Писемского

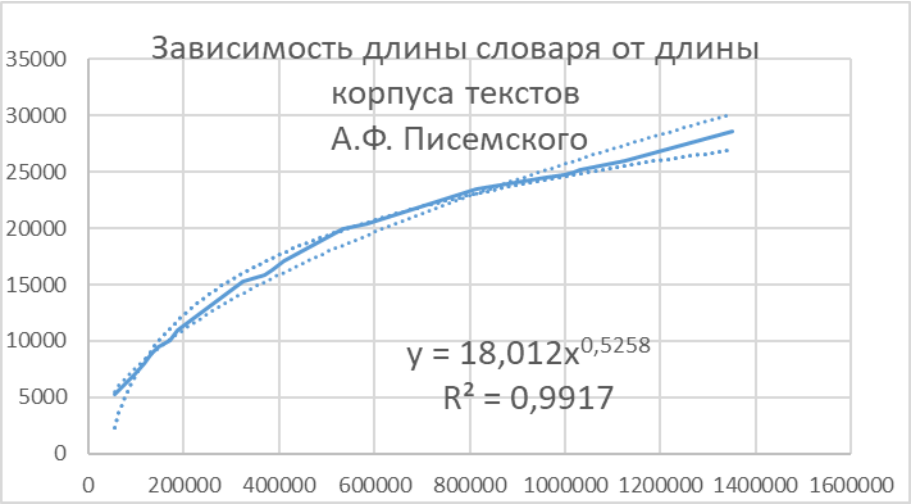

Рис. 3. Зависимость размера словаря от размера корпуса текстов А.Ф. Писемского

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):

$$N = 18.012 M^{0.5258}$$

Подставив в эту формулу значение $M_0 \approx e^{14,72273} \approx 2477418$, мы получим значение 41450 слов как оценку для предельного размера словаря. Это значение отличается от полученного ранее как нуля логарифмической функции тренда КЛР. При этом относительная погрешность составляет

$$\frac{48545 - 41450}{48545} \times 100\% \approx 14.61\%$$

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 А.Ф. Писемского находится в промежутке 41450 – 48545 слов,

- размер текста, при котором достигается предельный размер словаря А.Ф. Писемского, составляет 2.477.418 слов.

Теперь мы можем вычислить нижнюю оценку фрактальной размерности метакниги А.Ф. Писемского:

$$\alpha_0 \approx \frac{\ln 41450}{\ln 2477418} \approx 0.722166$$

Таким образом, фрактальная размерность метакниги А.Ф. Писемского, составленной из его 20 произведений, может быть заключена в промежуток [0,722166; 0,726969].

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 лексики / Ю.А. Тулдава. -— Таллин: Валгус, 1987. — 204 с.

10. Kretov A. A., Polovinkina M. V., Polovinkin I. P. and Lometc M.V. 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