Предельный размер словаря писателя и фрактальная размерность его метакниги

А. А. Кретов, email: kretov@rgph.vsu.ru [1]
М.В. Половинкина, email: polovinkina-marina@yandex.ru [2]
И. П. Половинкин, email: polovinkin@yandex.ru [1]
М. В. Ломец, email: marusya.lomets@gmail.com [1]

[1] Воронежский государственный университет,
[2] Воронежский государственный университет инженерных технологий

Аннотация. *В работе описываются возможности измерения таких характеристик моноавторских корпусов текстов, как предельный размер словаря писателя и фрактальная размерность его метакниги. Рассматривается проблема практического расчета фрактальной размерности. Приводятся результаты расчетов для метакниг Л.Н. Толстого и Ф.М. Достоевского.*

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

$$\Theta(M^{\alpha}) \text{ где } \alpha \in (0,1) \qquad (1)$$

Далее предлагается рассматривать закон Хипса не как асимптотическую оценку, а как точную формулу с переменным показателем α и переписать его в виде

$$\alpha = \alpha(M) = \ln N / \ln M . \qquad (2)$$

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

$$\alpha_0 = -\lim_{l \to 0} (\ln N / \ln l) = \lim_{l \to 0} (\ln N / \ln(1/l)) \qquad (3)$$

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

$$l = 1/M . \qquad (4)$$

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

$$\alpha_0 = -\lim_{l \to 0} (\ln N / \ln l) = \lim_{M \to +\infty} (\ln N / \ln M) = \lim_{M \to +\infty} \alpha(M), \quad (5)$$

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. Эксплуатируя концепцию метакниги, мы должны быть готовы отказаться от хронологической последовательности при рассмотрении совокупности произведений писателя. Важна лишь длина фрагмента метакниги. Однако не всегда речь идет о бессистемном перемешивании текстов входящих в метакнигу произведений. Предполагается, что в случае надобности возможна их конкатенация с соблюдением монотонности последовательности длин произведений, образующих метакнигу.

Авторы [5] утверждают и иллюстрируют примерами текстов трех разных авторов (Гарди, Мелвилла и Лоуренса), что α убывает с возрастанием M. Наши наблюдения за текстами Л.Н. Толстого этот вывод не опровергают. В этом смысле творчество Л.Н. Толстого представляет собой очень удобный объект исследования, поскольку Л.Н. Толстому принадлежат как сравнительно небольшие, так и весьма объемные («Война и мир») произведения, причем и диапазон между длинами маленьких и больших произведений заполнен достаточно плотно. Как известно, предел убывающей на промежутке функции в правом конце этого промежутка равен точной нижней грани функции на этом промежутке. Поэтому значение при максимальном значении в заданном диапазоне и следует считать наилучшим приближением верхней оценки фрактальной размерности.

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. Оценка фрактальной размерности метакниги Л.Н. Толстого

В качестве примера применения изложенных выше соображений мы рассмотрели 20 произведений Льва Толстого разного объема, охватывающие более-менее равномерно отрезок времени в 52 года. 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 Л.Н. Толстого, равной 0,7252. 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 таблице 1.

Таблица 1

Динамика КЛР в нарастающем корпусе текстов Л.Н. Толстого.

The dynamics of LD in the growing corpus of texts by L.N. Tolstoy.

№	Год	Текст	ΔN	ΔM	N	M	Y_{TTR}
1	1852	Детство	4253	30326	4253	30326	0,1402
2	1854	Отрочество	1452	23020	5705	53346	0,1069
3	1855	Севастопольские рассказы	2117	36041	7822	89387	0,0875
4	1856	Два гусара	844	17219	8666	106606	0,0813
5	1856	Утро помещика	751	15669	9417	122275	0,0770
6	1857	Юность	1208	49939	10625	172214	0,0617
7	1858	Альберт	147	7927	10772	180141	0,0598
8	1862	Поликушка	725	16879	11497	197020	0,0584
9	1863	Казаки	1391	46002	12888	243022	0,0530
10	1869	Война и мир	7212	459672	20100	702694	0,0286
11	1877	Анна Каренина	2702	270110	22802	972804	0,0234
12	1884	Записки сумасшедшего	32	3729	22834	976533	0,0234
13	1886	Смерть Ивана Ильича	190	17716	23024	994249	0,0232
14	1889	Крейцерова соната	270	25434	23294	1019683	0,0228

15	1890	Дьявол	395	14246	23689	1033929	0,0229
16	1891	Мать	48	3597	23737	1037526	0,0229
17	1895	Хозяин и работник	268	14270	24005	1051796	0,0228
18	1898	Отец Сергий	153	13706	24158	1065502	0,0227
19	1899	Воскресение	1591	137305	25749	1202807	0,0214
20	1904	Хаджи-Мурат	481	36376	26230	1239183	0,0212

В таблице 1 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.

Выберем в качестве линии тренда логарифмическую зависимость. Более точно, мы выбираем логарифмические и постоянные функции в качестве базисных функций в уравнении регрессии, а функцию зависимости КЛР от объема текста ищем в виде линейной комбинации базисных функций. Коэффициент правдоподобия в таком случае высок ($R^2 \approx 0,9611$). Мы получили уравнение вида

$$Y_{TTR} = 0,4081 - 0,028 \ln M \quad (6)$$

Значение размера текста в нуле этой функции мы можем считать соответствующим предельному размеру словаря. Эта функция достигает нулевого значения в точке $M_0 \approx 2129565$. Итак, исходя из выбранного способа моделирования, мы заключаем, что размер мета-книги, при котором достигается предельный размер словаря Л.Н. Толстого, составляет 2 129 565 слов. Ясно, что это некоторая приближенная оценка. Предельный объем словаря найдем из тех же соображений при том же выборе базисных функций. Мы получим уравнение

$$Y_{TTR} = 0,6301 - 0,0604 \ln M \quad (7)$$

Эта функция достигает нулевого значения в точке $N_0 \approx 33932$. Итак, оценка предельного размера N_0 словаря Л.Н. Толстого (с

необходимым замечанием об учете выбранного метода моделирования) составляет примерно 33 932 слова.

Есть еще одна проблема - проблема проверки достоверности полученных результатов. 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. Все же попробуем воспользоваться вариантом закона Ципфа

$$N = AM^{\gamma} \tag{8}$$

для описания зависимости размера словаря от размера текста. Это оправдано тем, что если из уравнений (5)--(6) исключить Y_{TTR}, получится зависимость вида (7). Воспользуемся первичными данными из таблицы 1 и аппроксимируем степенную функцию в законе Ципфа. Этот подход дает нам $N = 38,069\ M^{0,4653}$. Теперь, подставляя вместо M в эту формулу значение $M_0 \approx 2129565$, мы получим $N_0 \approx 33506$. Это значение отличается от полученного ранее как нуля логарифмической функции тренда КЛР. Однако относительная погрешность составляет

$$\frac{33932 - 33506}{33506} \times 100\ \% \approx 1,27\ \%,$$

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 Л.Н. Толстого составляет 33500 -- 34000 слов, размер текста, при котором достигается предельный размер словаря Л.Н. Толстого, составляет 2 129 500 – 2 130 000 слов. Теперь мы можем вычислить нижнюю оценку фрактальной размерности метакниги Л.Н. Толстого:

$$\alpha_0 \approx \frac{\ln 34000}{\ln 2130000} \approx 0,71605.$$

Таким образом, фрактальная размерность метакниги Л.Н. Толстого, составленной из его 20 книг, может быть заключена в промежуток [0,7160; 0,7252].

3. Оценка фрактальной размерности метакниги Ф.М. Достоевского

В качестве еще одного примера применения рассмотренного метода мы рассмотрели 17 произведений Ф.М. Достоевского и провели аналогичные действия. Мы пришли к верхней оценке фрактальной размерности метакниги Ф.М. Достоевского, равной 0,7190. Для нижней

оценки понадобилась фиксация всех промежуточных пар значений после каждой конкатенации. Эти данные приведены в таблице 2.

Таблица 2

Динамика КЛР в нарастающем корпусе текстов Ф.М. Достоевского.

The dynamics of LD in the growing corpus of texts by F.M. Dostoevskiy.

№	Год	Текст	ΔN	ΔM	N	M	Y_{TTR}
1	1846	Бедные люди	4798	42162	4798	42162	0,1138
2	1846	Двойник	2606	49342	7404	91504	0,0809
3	1847	Хозяйка	1260	23931	8664	115435	0,0751
4	1848	Белые ночи	447	17053	9111	132488	0,0688
5	1849	Неточка Незванова	1264	55290	10375	187778	0,0553
6	1859	Дядюшкин сон	1453	41538	11828	229316	0,0516
7	1859	Село Степанчиково и его обитатели	1693	66470	13521	295786	0,0457
8	1860	Записки из мертвого дома	2733	98476	16254	394262	0,0412
9	1861	Униженные и оскорблённые	1134	119692	17388	513954	0,0338
10	1862	Скверный анекдот	350	16414	17738	530368	0,0334
11	1864	Записки из подполья	611	35452	18349	531022	0,0346
12	1866	Игрок	537	44630	18886	575652	0,0328
13	1866	Преступление и наказание	2984	172635	21870	748287	0,0292
14	1869	Идиот	1657	209206	23527	957493	0,0246

15	1872	Бесы	1893	197329	25420	1154822	0,0220
16	1875	Подросток	1195	189590	26615	1344412	0,0198
17	1880	Братья Карамазовы	2822	297078	29437	1641490	0,0179

Выберем в качестве линии тренда снова логарифмическую зависимость. Мы получили уравнение вида

$$Y_{TTR} = 0{,}3603 - 0{,}025 \ln M \quad (9)$$

Значение размера текста в нуле этой функции мы можем считать соответствующим предельному размеру словаря. Эта функция достигает нулевого значения в точке $M_0 \approx 1815733$. Итак, исходя из выбранного способа моделирования, мы заключаем, что размер мета-книги, при котором достигается предельный размер словаря Ф.М. Достоевского, составляет 1 815 733 слов. Предельный объем словаря найдем из тех же соображений при том же выборе базисных функций. Мы получим уравнение

$$Y_{TTR} = 0{,}5283 - 0{,}05 \ln M \quad (10)$$

Эта функция достигает нулевого значения в точке $N_0 \approx 38793$. Итак, оценка предельного размера N_0 словаря Ф.М. Достоевского (с необходимым замечанием об учете выбранного метода моделирования) составляет примерно 38 793 слова.

Воспользуемся первичными данными из таблицы 2 и аппроксимируем степенную функцию в законе Ципфа. Этот подход дает нам $N = 27{,}484\, M^{0{,}4908}$. Теперь, подставляя вместо M в эту формулу значение $M_0 \approx 1815733$, мы получим $N_0 \approx 32435$. Это значение отличается от полученного ранее как нуля логарифмической функции тренда КЛР. Относительная погрешность составляет

$$\frac{38793 - 32435}{32435} \times 100\ \% \approx 19{,}60\ \%.$$

На сей раз относительная погрешность весьма велика. Поэтому нижняя оценка фрактальной размерности будет не очень точна. Придется выбрать вариант, который дает меньшее значение α_0:

$$\alpha_0 \approx \frac{\ln 32435}{\ln 1815733} \approx 0{,}72071839\ 5027795851\ 7054339469\ 5242.$$

Мы столкнулись с неожиданным эффектом: значение величины, которой мы отводили роль нижней оценки фрактальной размерности, оказалось больше значения, которое мы считали верхней оценкой. Поэтому мы вынуждены лишь констатировать, что фрактальная размерность метакниги Ф.М. Достоевского не превосходит 0,7191.

Литература

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. Бронник Л. В. О фрактальном самоподобии в языке / Л. В. Бронник // Известия Волгоградского государственного педагогического университета. – 2009. – Т. 44, № 10. – С. 15-19.
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.