УДК 519.22

О ПАРАМЕТРИЧЕСКОЙ МОДЕЛИ РАСПРЕДЕЛЕНИЯ ДЛИНЫ СЛОВ НА ПРИМЕРЕ ЛИТЕРАТУРНЫХ ТЕКСТОВ НА ИСПАНСКОМ, ИТАЛЬЯНСКОМ И ШВЕДСКОМ ЯЗЫКАХ

ON THE PARAMETRIC MODEL OF LENGTH DISTRIBUTION OF THE WORDS ON THE LITERARY TEXTS EXAMPLE IN SPANISH, ITALIAN AND SWEDISH LANGUAGES

©Палий И. А.
Сибирский государственный автомобильно-дорожный университет
Россия, г. Омск, paliy_ia@mail.ru
©Palii I.
Siberian State Automobile and Highway University
Russia, Omsk, paliy_ia@mail.ru

Аннотация. Исследуются закономерности, которым подчиняются относительные частоты длин слов, если разбить весь ряд относительных частот на несколько отрезков.

В случае испанского языка отрезков четыре: длины 1-2 (линейная функция $y = a_0 + a_1 n$ с положительным наклоном); длины 3-5 (полином второго порядка $y = a_0 + a_1 n + a_2 n^2$ с ветвями, направленными вверх); длины 6-11 (линейная функция с отрицательным наклоном); длины 12 и более (геометрическая прогрессия $y = aq^n$ со знаменателем меньше 1). Здесь n – длина слова (число букв в нем).

В случае итальянского языка отрезков тоже четыре: длины 1-3 и 4-6 (полиномы второго порядка с ветвями, направленными вниз); длины 7-11 (геометрическая прогрессия со знаменателем меньше 1); длины 12 и более (геометрическая прогрессия со знаменателем меньше 1).

В случае шведского языка отрезков три: длины 1-3 (полином второго порядка с ветвями, направленными вверх); длины 4-6 (полином второго порядка с ветвями, направленными вниз); длины 7 и более (геометрическая прогрессия со знаменателем меньше 1).

Коэффициенты уравнений – это параметры, которые можно оценить для данного текста на основании его статистических характеристик.

Рассматривались пять текстов на испанском и шведском языках и шесть текстов на итальянском языке. Затем все тексты на данном языке объединялись в один текст и рассматривалось распределение относительных частот длин слов в таком объединенном тексте.

Abstract. We study regularities, to which the relative frequencies of the word lengths are subject, if the entire series of relative frequencies is divided into several segments.

In the case of the Spanish language, there are four segments: lengths 1-2 (linear function $y = a_0 + a_1 n$ with positive slope); Lengths 3-5 (a polynomial of the second order $y = a_0 + a_1 n + a_2 n^2$ with branches directed upwards); Lengths 6-11 (linear function with negative slope); Length 12 and more (geometric progression $y = aq^n$ with a denominator less than 1). Here n is the length of the word (the number of letters in it).

In the case of the Italian language, there are also four lengths: lengths 1-3 and 4-6 (polynomials of the second order with branches directed downwards); Length 7-11 (geometric progression with

denominator less than 1); Length 12 and more (geometric progression with a denominator less than 1).

In the case of the Swedish language, there are three segments: lengths 1-3 (a second-order polynomial with branches pointing upwards); Length 4-6 (second-order polynomial with branches directed downwards); Length 7 and more (geometric progression with a denominator less than 1).

Coefficients of equations are parameters that can be estimated for a given text on the basis of its statistical characteristics.

Five texts in Spanish and Swedish and six texts in Italian were considered. Then all the texts in the given language were combined into one text and distribution was considered.

Ключевые слова: текст на испанском языке, текст на итальянском языке, текст на шведском языке, длины слов, параметрическая модель распределения длины слов.

Keywords: text in Spanish, text in Italian, text in Swedish, word length, parametric model of word-length distribution.

Мы продолжаем начатое в работах [1–2] исследование закономерностей, позволяющих описать распределение относительных частот длин слов в данном языке при помощи нескольких функциональных зависимостей.

Исследовались литературные тексты на испанском, итальянском и шведском языках. В тексты были внесены некоторые изменения. Апостроф в текстах на итальянском языке считался буквой. Кроме того, из текстов были, по возможности, удалены инициалы и сокращения и некоторые другие посторонние включения.

Список использованных текстов приведен в Таблице 1. Все тексты представлены на сайте gutenberg.org.

Относительные частоты длин слов текстов 1-8 приведены в табл. 2; текстов 9-16 – в Таблице 3. Максимальная длина слов, указанная в Таблицах 2-3, равна 25, но в тексте «Человек без юмора» есть одно слово длины 28, оно в табл. 3 не вошло. Относительные частоты указаны с пятью знаками после запятой.

Таблица 1

ИСПОЛЬЗОВАННЫЕ ТЕКСТЫ

№	*Язык*	*Автор*	*Название*	*Число слов*
1	Испанский	Х.К. Давалос	Сальта	29832
2	Испанский	М. де Сервантес	Дон Кихот	376491
3	Испанский	В. Бласко Ибаньес	Кровь и песок	111357
4	Испанский	Л. Алас и Уренья	Единственный сын	89347
5	Испанский	Э. Хоуп	Узник Зенды (перевод на испанский язык)	50959
6	Итальянский	А. Альбертацци	Старинные истории о любви	28603
7	Итальянский	А. Мандзони	Обрученные	214861
8	Итальянский	А. Виванти	Поглотители	103217
9	Итальянский	А. Баррили	Ночь командора	108042
10	Итальянский	Р. Джероламо	Мать скорбящая	125966
11	Итальянский	И. Баччини	Рассказы для детей	30783
12	Шведский	А. Энгстрем	Моя жизнь и времена	39558
13	Шведский	Д. Андерссон	Наследие Давида Рамм	52148
14	Шведский	Г. Гейерстам	Книга о маленьком братце	47399
15	Шведский	Э.Г. Хельстрем	Человек без юмора, т.1	93387
16	Шведский	С. Лагерлеф	Изгнанник	76063

Таблица 2

ОТНОСИТЕЛЬНЫЕ ЧАСТОТЫ ДЛИН СЛОВ В ТЕКСТАХ 1-8

Длина слова	*Текст 1*	*Текст 2*	*Текст 3*	*Текст 4*	*Текст 5*	*Текст 6*	*Текст 7*	*Текст 8*
1	0,05833	0,07695	0,05594	0,06495	0,06548	0,10684	0,07175	0,07476
2	0,25637	0,23421	0,24402	0,24662	0,24751	0,17135	0,17236	0,17125
3	0,14401	0,17421	0,15495	0,15109	0,14096	0,14740	0,16452	0,14227
4	0,08115	0,09645	0,08582	0,09305	0,09708	0,08569	0,08619	0,09658
5	0,11219	0,12028	0,10897	0,12404	0,12306	0,13624	0,14376	0,16604
6	0,09553	0,09647	0,09161	0,08953	0,08878	0,10317	0,10903	0,10478
7	0,08474	0,07854	0,08452	0,07579	0,09013	0,09038	0,08493	0,09182
8	0,06587	0,05268	0,06860	0,05625	0,06176	0,06569	0,05978	0,05889
9	0,04029	0,03431	0,04222	0,04434	0,03713	0,04129	0,03939	0,03680
10	0,02903	0,01934	0,02922	0,02674	0,02198	0,02765	0,03107	0,02620
11	0,01606	0,00845	0,01753	0,01357	0,01277	0,01339	0,01782	0,01377
12	0,00962	0,00457	0,00910	0,00685	0,00665	0,00675	0,00950	0,00860
13	0,00379	0,00221	0,00419	0,00395	0,00347	0,00255	0,00513	0,00422
14	0,00198	0,00090	0,00213	0,00207	0,00224	0,00119	0,00249	0,00220
15	0,00070	0,00029	0,00075	0,00094	0,00075	0,00028	0,00147	0,00109
16	0,00027	0,00011	0,00037	0,00015	0,00016	0,00010	0,00048	0,00041
17	0,00003	0,00002	0,00005	0,00003	0,00004	0,00003	0,00017	0,00020
18	0,00003	5,31E-06	0	0,00003	0,00006	-	0,00010	0,00009
19	-	5,31E-06	8,98E-06	-	-	-	0,00003	0,00003
20	-	0	-	-	-	-	0,00001	9,68E-06
21	-	2,66E-06	-	-	-	-	-	

Таблица 3

ОТНОСИТЕЛЬНЫЕ ЧАСТОТЫ ДЛИН СЛОВ В ТЕКСТАХ 9-16

Длина слова	*Текст 9*	*Текст 10*	*Текст 11*	*Текст 12*	*Текст 13*	*Текст 14*	*Текст 15*	*Текст 16*
1	0,06316	0,06120	0,07693	0,02945	0,02142	0,01658	0,02623	0,01742
2	0,18708	0,17741	0,19121	0,13906	0,13308	0,10726	0,13218	0,12079
3	0,16109	0,15103	0,13475	0,28960	0,33296	0,35497	0,30588	0,34929
4	0,08716	0,07734	0,08917	0,12020	0,13866	0,15513	0,14350	0,15750
5	0,13317	0,15002	0,14888	0,12193	0,12710	0,14694	0,12894	0,12687
6	0,10457	0,10954	0,10860	0,09978	0,09180	0,10268	0,09173	0,09546
7	0,08973	0,09330	0,08826	0,06250	0,05613	0,04243	0,06227	0,04450
8	0,06357	0,06233	0,05860	0,04724	0,03829	0,02838	0,03691	0,02980
9	0,04184	0,04394	0,04132	0,03029	0,02479	0,01985	0,02609	0,02411
10	0,03214	0,03463	0,02920	0,02230	0,01607	0,01247	0,01706	0,01362
11	0,01842	0,01912	0,01530	0,01287	0,00886	0,00650	0,01041	0,00920
12	0,00938	0,01022	0,00926	0,01037	0,00462	0,00340	0,00708	0,00572
13	0,00478	0,00490	0,00455	0,00580	0,00265	0,00152	0,00436	0,00256
14	0,00251	0,00292	0,00240	0,00345	0,00155	0,00089	0,00289	0,00116
15	0,00094	0,00139	0,00088	0,00218	0,00067	0,00057	0,00188	0,00101
16	0,00026	0,00039	0,00055	0,00124	0,00058	0,00021	0,00086	0,00039
17	0,00013	0,00025	0,00010	0,00048	0,00023	0,00017	0,00054	0,00030
18	0,00004	0,00003	0,00003	0,00053	0,00021	0,00002	0,00036	0,00021
19	0,00003	0,00002	-	0,00035	0,00013	0,00002	0,00048	0,00001
20	0	0,00003	-	0,00015	0,00010	0	0,00014	0,00001
21	0	-	-	0,00005	0,00004	0	0,00011	0,00001
22	0,00003	-	-	0,00008	0,00004	0,00002	0,00007	0,00001
23	-	-	-	0,00005	0	-	0,00001	0
24	-	-	-	0,00005	0	-	0	0,00001
25	-	-	-	-	0,00002	-	0,00002	

На Рисунках 1-16 представлены распределения частот длин слов текстов 1-16. Массивы длин слов разбиты на несколько отрезков. Для каждого отрезка построен свой тренд, соответствующий распределению относительных частот длин слов на этом отрезке, приведены уравнение тренда и значение коэффициента детерминации, автоматически рассчитанные MS EXCEL. На Рисунках 17-19 представлены те же данные, когда пять текстов на испанском языке, 6 текстов на итальянском языке, 5 текстов на шведском языке рассматривались как один текст (тексты 17, 18, 19 соответственно).

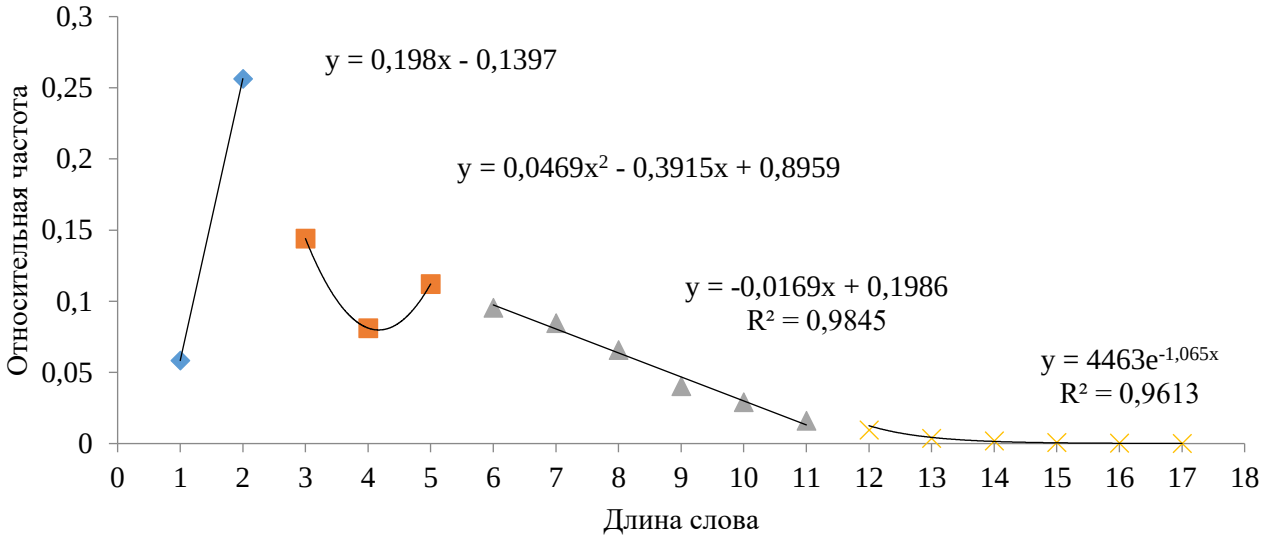

Рисунок 1. Х.К. Давалос, Сальта. Относительные частоты длин слов

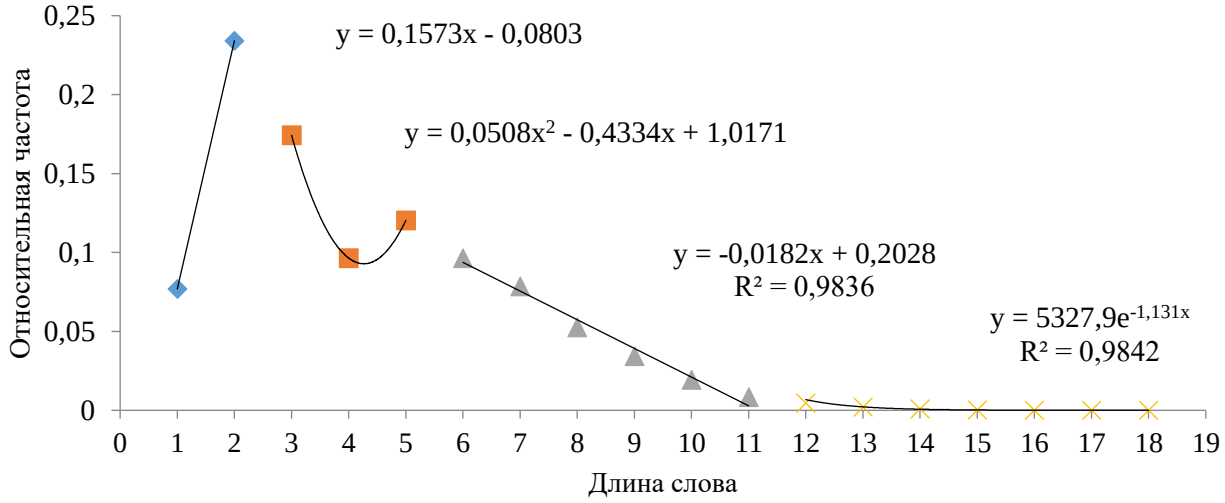

Рисунок 2. М. де Сервантес, Дон Кихот. Относительные частоты длин слов

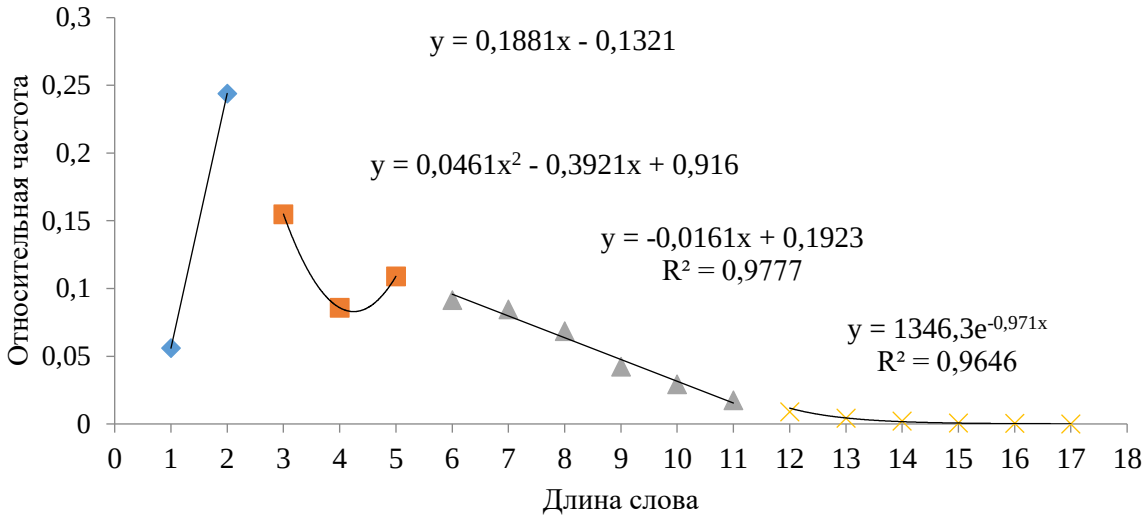

Рисунок 3. В. Бласко Ибаньес, Кровь и песок. Относительные частоты длин слов

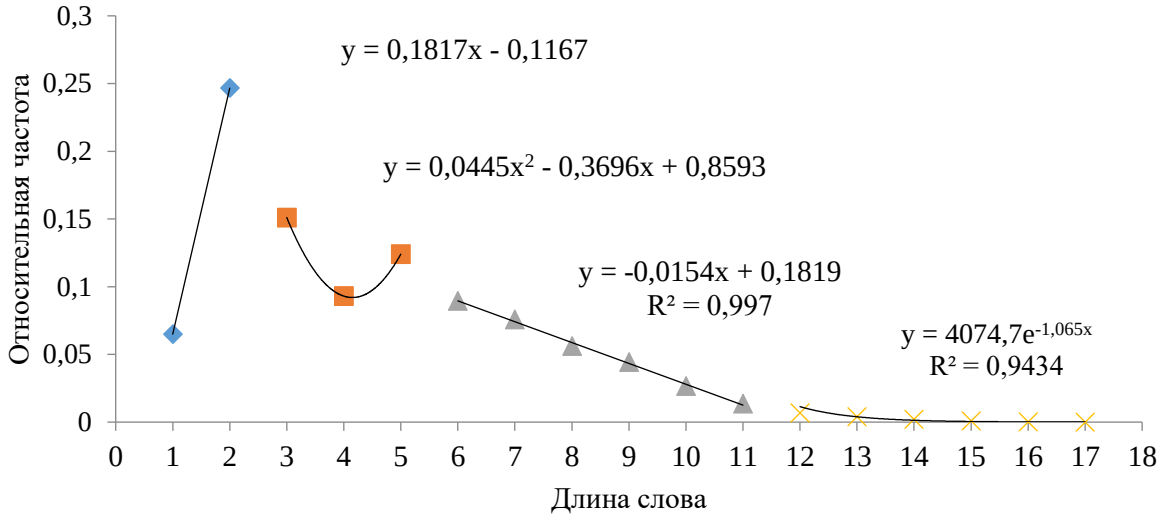

Рисунок 4. Л. Алас и Уренья, Единственный сын. Относительные частоты длин слов

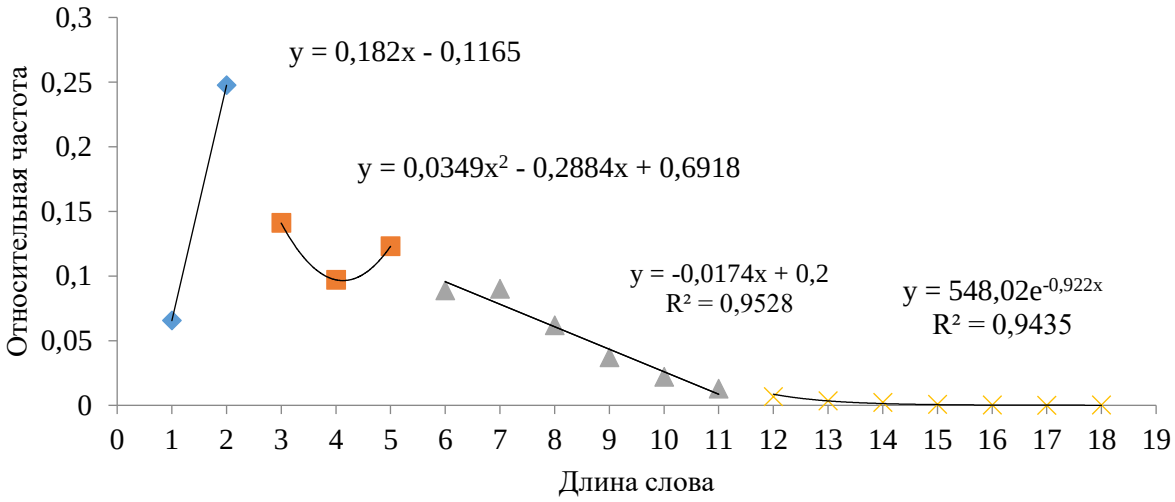

Рисунок 5. Э. Хоуп, Узник Зенды (перевод на испанский). Относительные частоты дли слов

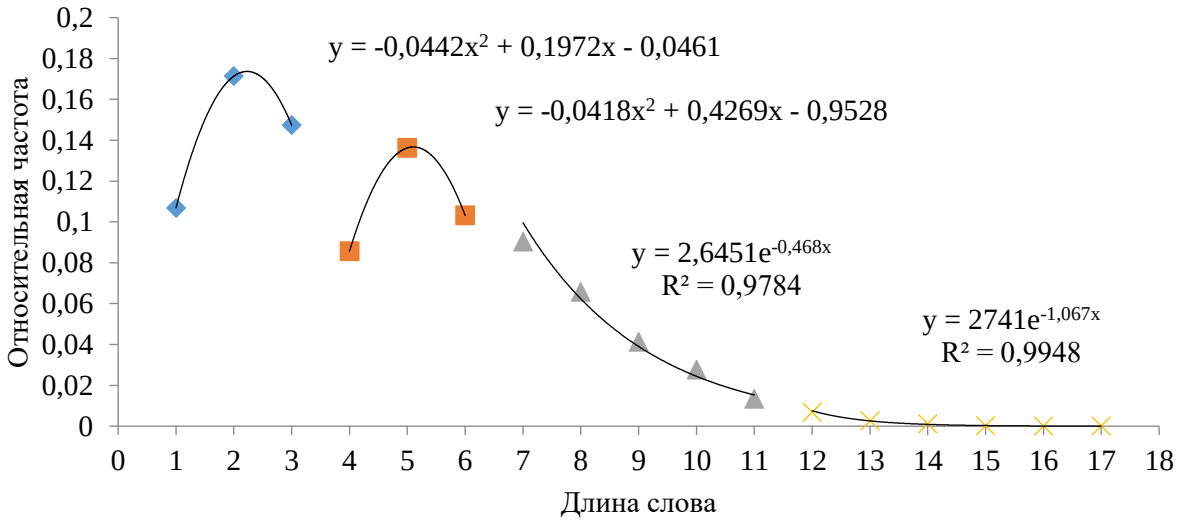

Рисунок 6. А. Альбертацци, Старинные истории о любви. Относительные частоты длин слов

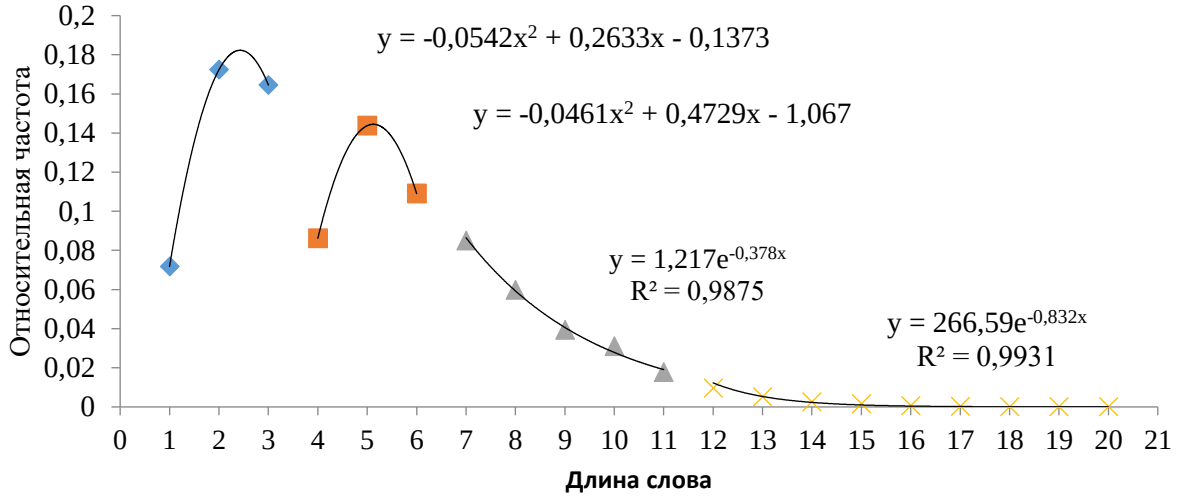

Рисунок 7. А. Мандзони, Обрученные. Относительные частоты длин слов

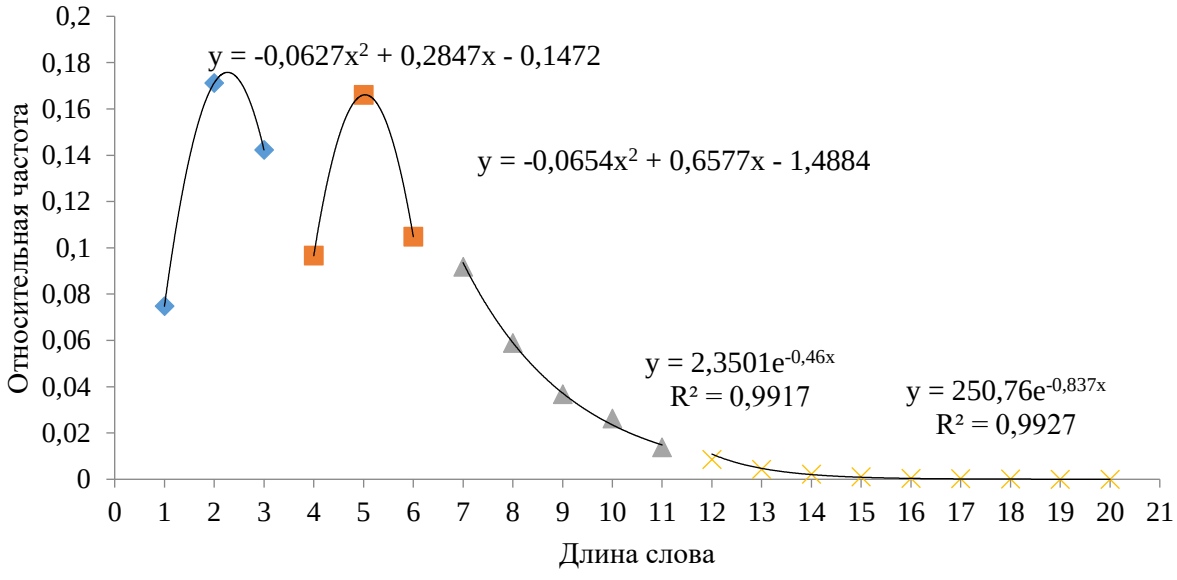

Рисунок 8. А. Виванти, Поглотители Относительные частоты длин слов

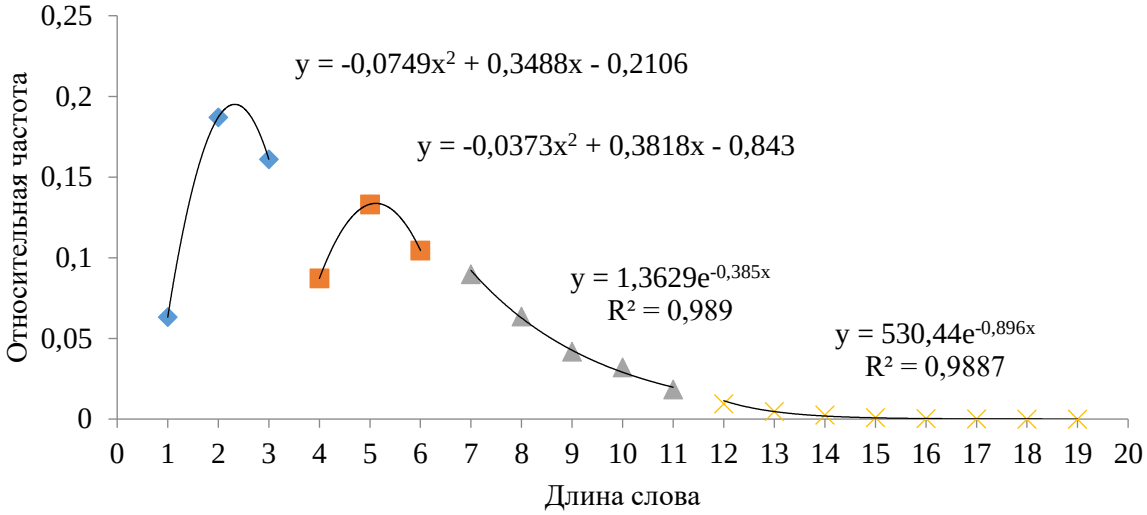

Рисунок 9. А. Баррили, Ночь командора. Относительные частоты длин слов

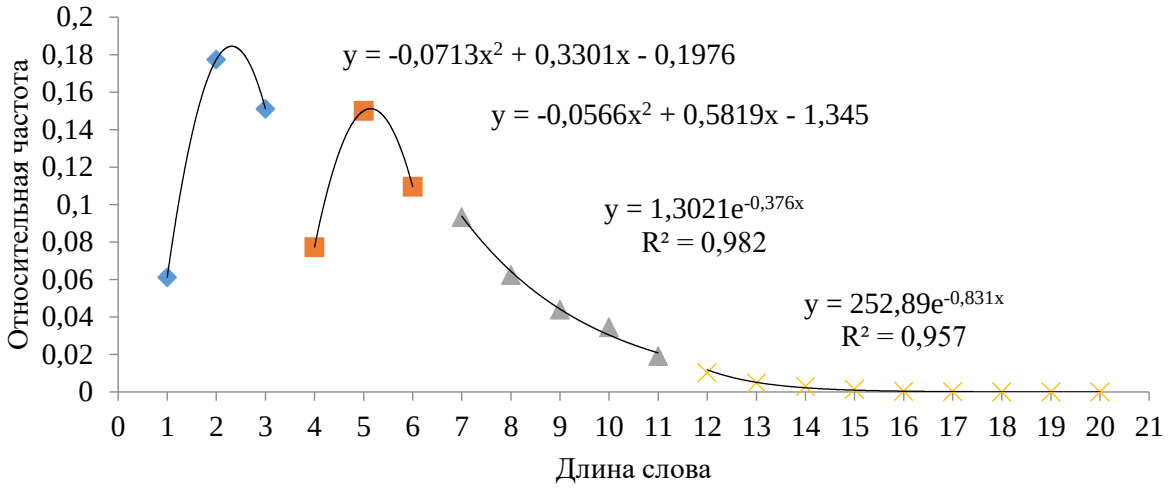

Рисунок 10. Р. Джероламо, Мать скорбящая. Относительные частоты длин слов

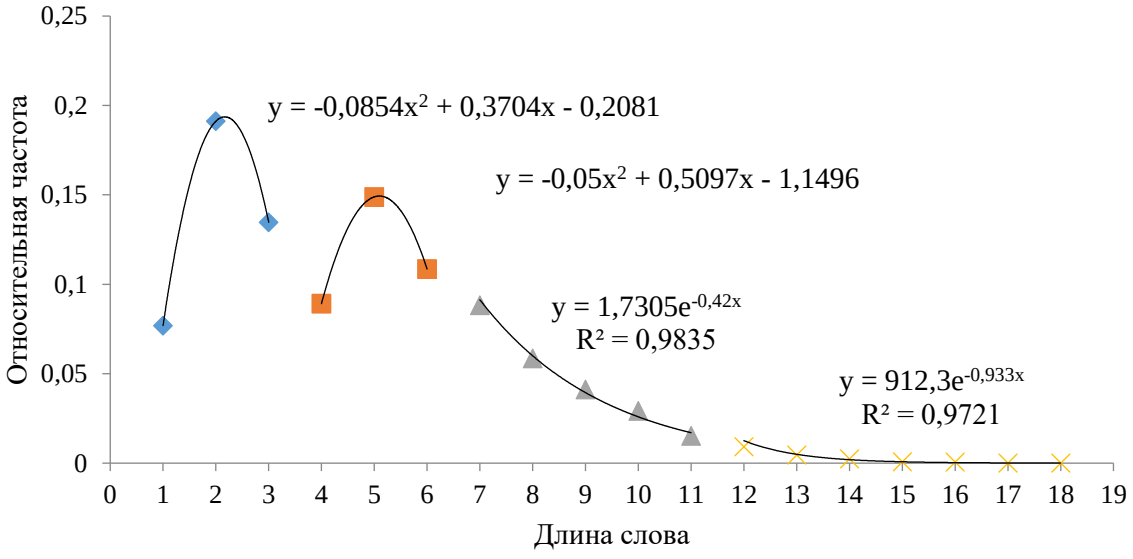

Рисунок 11. И. Баччини, Рассказы для детей. Относительные частоты длин слов

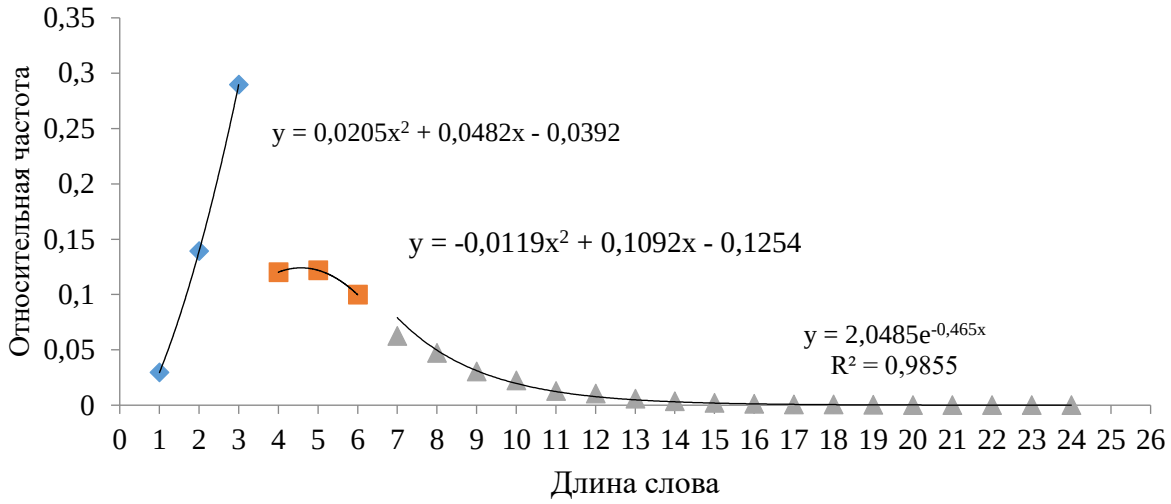

Рисунок 12. А. Энгстрем, Моя жизнь и времена. Относительные частоты длин слов

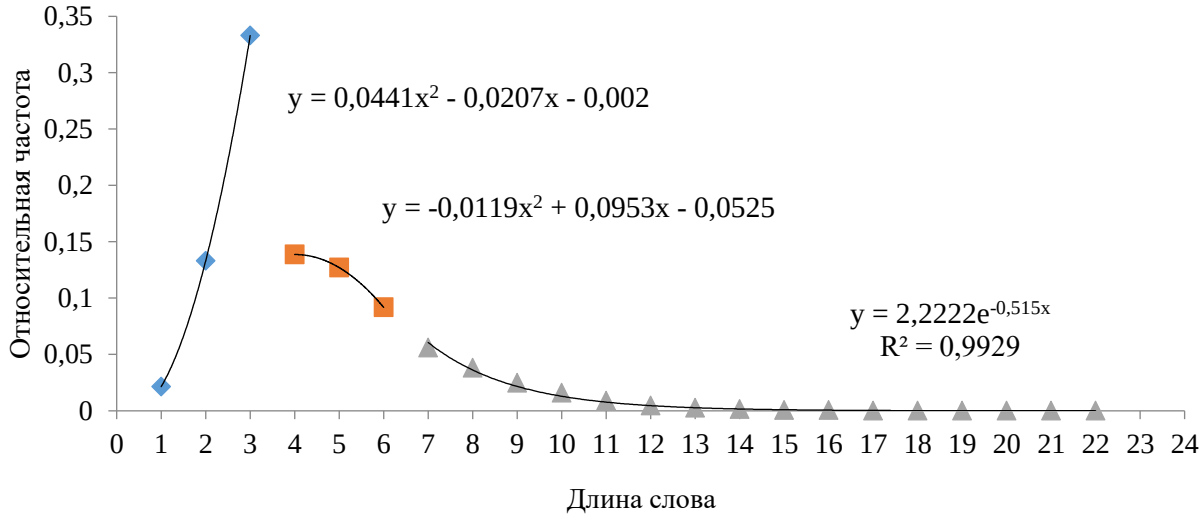

Рисунок 13. Д. Андерссон, Наследие Давида Рамм.Относительные частоты длин слов

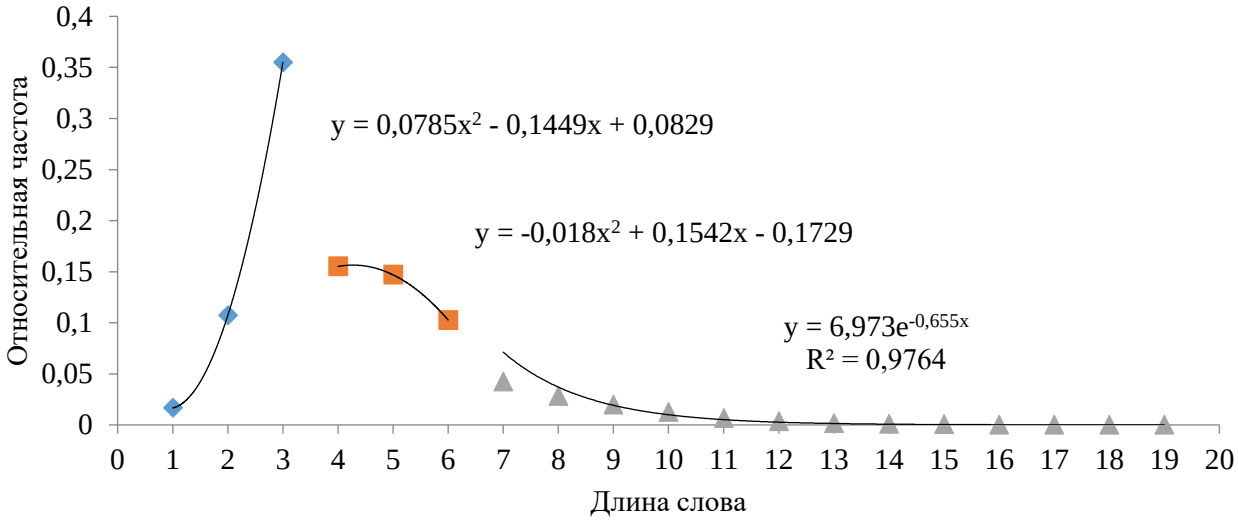

Рисунок 14. Г. Гейерстам, Книга о маленьком братце. Относительные частоты длин слов

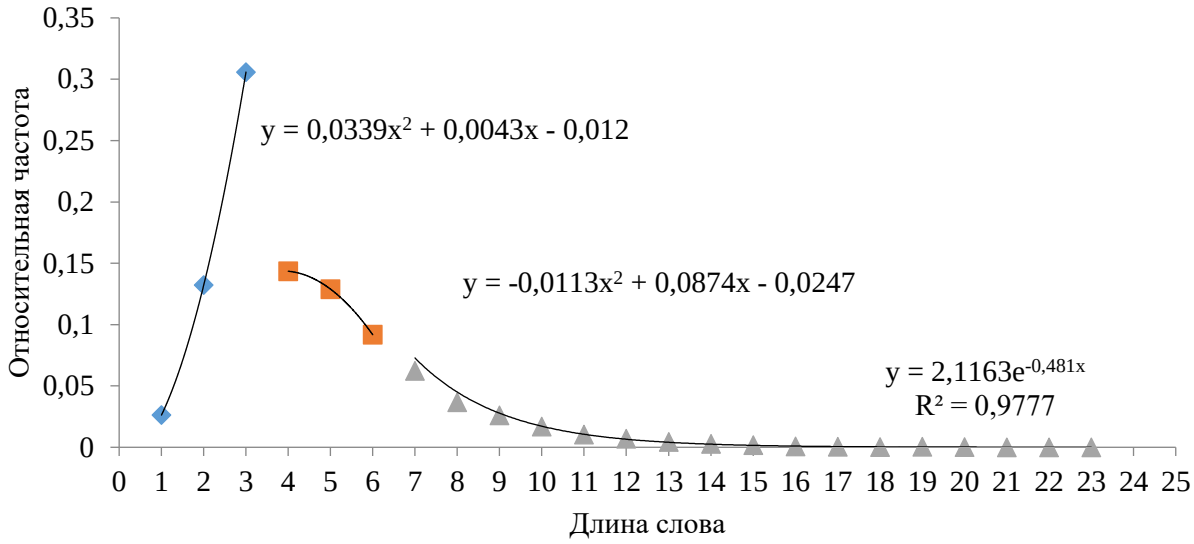

Рисунок 15. Э. Г. Хельстрем, Человек без юмора. Относительные частоты длин слов

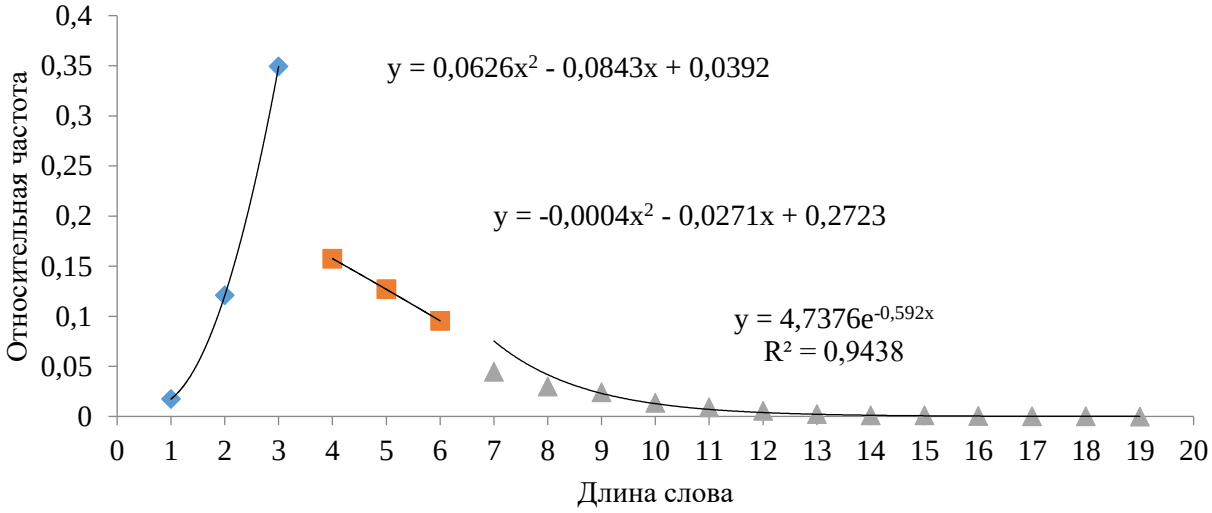

Рисунок 16. С. Лагерлеф, Изгнанник. Относительные частоты длин слов

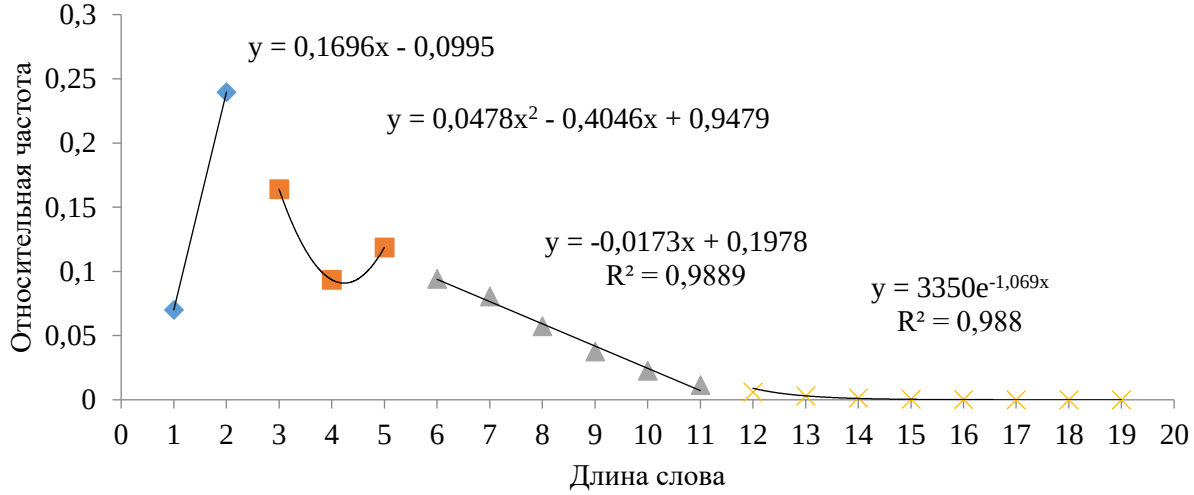

Рисунок 17. Испанский язык, объединение 5 текстов. Относительные частоты длин слов

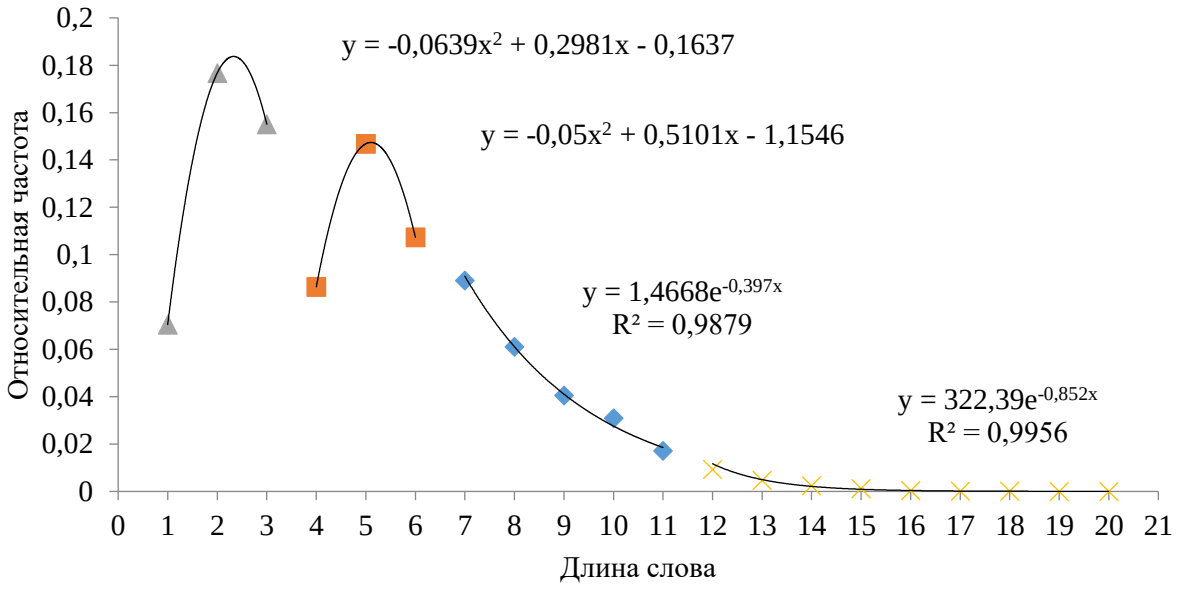

Рисунок 18. Итальянский язык, объединение 6 текстов. Относительные частоты длин слов

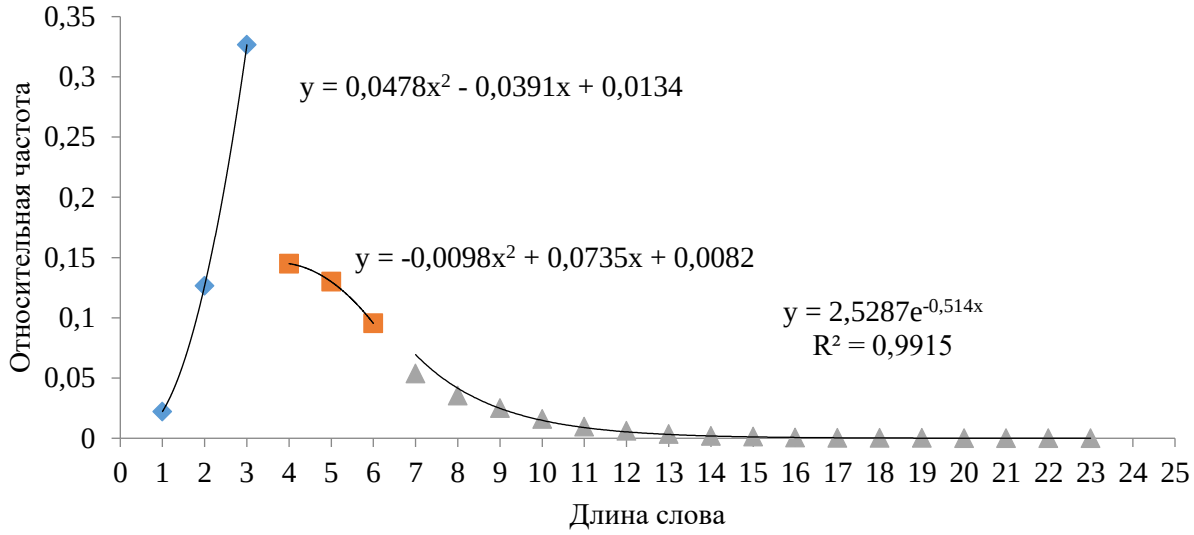

Рисунок 19. Шведский язык, объединение 5 текстов. Относительные частоты длин слов

Пусть весь массив длин слов разбит на k отрезков $[n_i^1, n_i^2], i = 1, \cdots, k$;

$f_i -$ функциональная зависимость, по которой вычисляются относительные частоты длин, принадлежащих отрезку с номером i.

Тогда должно выполняться равенство

$$\sum_{i=1}^{k} \sum_{n=n_i^1}^{n_i^2} f_i(n) = 1 \quad (1)$$

Коэффициенты функциональных зависимостей подбирались следующим образом.

Полиномами второго порядка $f(n) = a_0 + a_1 n + a_2 n^2$ описывалось поведение частот на отрезках длин слов, включающих три длины. Три коэффициента этих полиномов однозначно вычисляются по экспериментальным данным. Так же однозначно вычисляются два коэффициента линейного уравнения, если нужно соединить отрезком две экспериментальные точки.

Коэффициенты линейного уравнения $f(n) = a_0 + a_1 n$, когда число экспериментальных точек больше двух, находились по методу наименьших квадратов.

Коэффициенты уравнения $f(n) = ae^{bn}$ в случае отрезка длин слов, не являющегося последним, вычислялись по методу наименьших квадратов.

Во всех рассмотренных текстах относительные частоты длин слов, начиная с длины 12 (длины 7 для шведского языка), убывают примерно по геометрической прогрессии.

Коэффициенты геометрической прогрессии $f(n) = aq^n$ определялись для объединенных текстов 17, 18, 19 следующим образом. Сначала подбиралось значение $b = aq^{12}$ (aq^7 для шведского языка). Это значение подбиралось так, чтобы минимизировать сумму квадратов отклонений экспериментальных относительных частот от частот, определяемых геометрической прогрессией $y = aq^n$. Именно подбиралось, потому что вывести формулу подсчета коэффициента b, как это делается по методу наименьших квадратов для коэффициентов линейной функции, не представляется возможным. Затем, исходя из условия (1), вычислялось значение $c = \sum_{n=12}^{\infty} aq^n = \frac{aq^{12}}{1-q} = b/(1-q)$.

Тогда $q = 1 - \frac{b}{c}, a = \frac{b}{q^{12}}$.

В Таблице 4 указаны найденные значения a и q для текстов 17-19 и величины коэффициентов детерминации (R^2). В каждом из трех случаев значения коэффициентов детерминации почти не отличаются от 1, значения экспериментальных относительных частот длин слов близки значениям, которые задаются формулой $y = aq^n$. Относительные частоты длин слов, когда $n > 11$ малы (тысячные, десятитысячные), поэтому особенно подвержены влиянию случайности. Тем не менее, все значения R^2 больше 0,99.

Таблица 4

АППРОКСИМАЦИЯ ЗНАЧЕНИЙ ОТНОСИТЕЛЬНЫХ ЧАСТОТ ДЛИН СЛОВ ГЕОМЕТРИЧЕСКОЙ ПРОГРЕССИЕЙ

№ текста	*Длина слова n*	*Язык*	*a*	*q*	R^2
17	$n \geq 12$	Испанский	92,16	0,448	0,9959
18	$n \geq 12$	Итальянский	80,92	0,469	0,9948
19	$n \geq 7$	Шведский	1,146	0,646	0,9977

Некоторые выводы.

Представляется вероятным, что выявленные закономерности поведения относительных частот длин слов не являются случайными и присущи в той или иной степени всем литературным текстам на данном языке. Чтобы выводы стали определеннее, нужно исследовать даже не десятки, а сотни литературных текстов.

Программы, позволившие провести данное исследование, написаны студентами факультета "Информационные системы управления" СибАДИ М. С. Петровой и С. В. Суторминым.

Список литературы:

1. Палий И. А. О параметрической модели распределения длины слов на примере языка иврит // Science and World. International Scientific Journal. 2017. № 1 (41), Т 1. С. 8-11.
2. Палий И. А. О параметрической модели распределения длины слов на примере литературных текстов на немецком, французском и новогреческом языках // Science and World. International Scientific Journal. 2017. № 3 (43), Т 1. С. 24-29.

References:

1. Paliy, I. (2017). On the parametric model of word-length distribution on the example of Hebrew language. Science and World. *International Scientific Journal,* 1(41), T. 1, 8-11.

2. Paliy, I. (2017). On the parametric model of word-length distribution on the example of literary texts in German, French and Modern Greek languages. *Science and World. International Scientific Journal,* 3 (43), T. 1, 24-29.

Работа поступила
в редакцию 01.08.2017 г.

Принята к публикации
04.08.2017 г.

Ссылка для цитирования:

Палий И. А. О параметрической модели распределения длины слов на примере литературных текстов на испанском итальянском и шведском языках // Бюллетень науки и практики. Электрон. журн. 2017. №8 (21). С. 10-21. Режим доступа: http://www.bulletennauki.com/palii (дата обращения 15.08.2017).

Cite as (APA):

Palii, I. (2017). On the parametric model of length distribution of the words on the literary texts example in spanish italian and swedish languages. *Bulletin of Science and Practice*, (8), 10-21